

Fast text-to-speech algorithms for Esperanto, Spanish, Italian, Russian and English

BRUCE ARNE SHERWOOD

*Computer-based Education Research Laboratory and Department of Physics,
University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, U.S.A.*

(Received 28 November 1977, and in revised form 10 June 1978)

Fast, simple algorithms are presented for producing good-quality speech from a phonemic synthesizer in five different languages: Esperanto, Spanish, Italian, Russian and English. The Esperanto and Spanish algorithms can operate on standard written texts. For Italian the typist must occasionally mark the stressed syllable in a word, and for Russian the typist must mark all stressed syllables. For English, the typist must use a phonetic spelling and mark the stressed syllables: situations where this approach is viable include the addition of speech output to computer-based educational materials. A particularly advantageous phonetic spelling scheme for English is presented. Comparisons are made among the five languages as to the ease with which text-to-speech conversions can be made.

Introduction

The purpose of this article is to describe simple algorithms for driving a phonemic speech synthesizer to produce relative high-quality speech from stored or computer-generated text in Esperanto, Spanish, Italian, Russian or English. The particular synthesizer used was the Votrax model VS-6,[†] but the techniques apply to any phonemic synthesizer (that is, a synthesizer which produces speech from an input string of phoneme codes, as opposed to synthesizers whose inputs are vocal-tract parameters such as resonance frequencies).

Phonemic synthesizers are attractive speech output devices in computer systems (Witten & Madams, 1977[‡]). The text which feeds such synthesizers requires very little storage space and very low transmission bandwidth (a speaking rate of 120 words per minute requires only about 80 bits per second of phoneme transmission). Moreover, techniques for generating and manipulating text are well-developed in computer

[†] The Votrax VS-6 is made by Federal Screw Works, 500 Stephenson Highway, Troy, Michigan 48084, who also make a more elaborate model, the ML-1. A version of the VS-6 without programmable pitch variations is marketed by The Digital Group, P.O. Box 6528, Denver, Colorado 80206. A less powerful phonemic synthesizer is the model 1000 of AI Cybernetic Systems, P.O. Box 4691, University Park, New Mexico 88003. These seem to be the only phonemic synthesizers presently on the market. An equivalent device could in principle be constructed by combining a vocal-parameter (format) synthesizer with a microprocessor whose program would convert from text to synthesizer parameters. Such a combination, including 8080-compatible hardware and software, is available commercially from Computalker Consultants, P.O. Box 1951, Santa Monica, California 90406. At the present time the Votrax device speaks significantly better than any of these other devices but not as well as some research laboratory speech-synthesis systems.

[‡] Witten & Madams (1977) actually used a vocal-parameter synthesizer, not a phonemic synthesizer. However, they generated synthetic speech by program from stored (phonetic) text in real time, and the functional equivalence of their system to those which use phonemic synthesizers makes many of their comments relevant to such systems.

systems. In comparison, compressed natural speech currently requires ten to twenty times as much data to store and transmit (Schafer & Rabiner, 1975; Makhoul, 1975), and associated editing and generation techniques are less well developed. Recorded speech from serial-access or random-access tape or disk systems (Yeager, 1976) requires very little bandwidth to choose the appropriate message, but distribution and updating can be difficult and it is usually impossible by program to assemble a natural-sounding sentence from component words. However, both compressed and recorded speech can be of much higher quality than speech produced by a phonemic synthesizer, at least at present.

In the case of English, complex and irregular spelling and sound structures have made it difficult to use phonemic synthesizers. In the broad spectrum of attacks on these problems of English, it is useful to identify two extreme approaches. One is to have a short list of highly irregular English words plus a set of rules for attempting to interpret all other words. Generally this results in a certain percentage of words being mispronounced, but much text is nevertheless understandable in context (McIlroy, 1974; Elovitz, Johnson, McHugh & Shore, 1976). Such algorithms can run at speech rates in a microcomputer, so that a talking computer terminal for the blind becomes feasible (Kutsch, 1976[†]). Coupling such a device to a character-recognition machine can make it possible for a blind person to read a newspaper.[‡] In such applications even relatively low-quality speech is extremely useful. The other extreme is well represented by a project to produce high-quality readings of books for the blind from publishers' tapes of machine-readable text resulting from computer typesetting (Allen, 1976). Here a large dictionary of English words and roots, coupled with sophisticated spelling algorithms (for rare forms not found in the dictionary), and parsing algorithms (for phrasing) can produce high-quality speech. Such procedures currently require many seconds of processing by rather powerful computers to produce one second of speech.

There are situations where neither of these extremes is appropriate—where high-quality speech is required but where little processing is feasible. An example of such situations is the problem of adding some speech output to computer-based educational materials. In general, if the text is *new* text rather than *existing* text, some sharing of the processing load with the human author is feasible. If the author enters the new text in some kind of phonetic alphabet, with the stressed syllables marked, many of the really hard problems are bypassed, and reasonably high-quality English can be produced at low computer cost. The challenge is to devise a spelling scheme which is acceptable to the human author in terms of ease of use. One such scheme will be described later.

On the other hand, some languages are written phonetically. As will be discussed in detail in this article, Esperanto and Spanish have such simple and accurate written forms that it is easy to produce good-quality speech from standard text. In both languages, stress locations as well as sounds can be reliably deduced from the text. Italian and Russian are also easier to handle than English, though it is necessary to mark stressed syllables in these languages, since there are no simple rules for determining stress location.

[†] Peter Maggs of the University of Illinois Law School has recently written an 8080-microcomputer implementation of the algorithm of Elovitz *et al.*, in order to provide computer services to the blind in the form suggested by Kutsch. This algorithm has certain advantages over the McIlroy algorithm used by Kutsch.

[‡] Such a reading machine is now being produced by Kurzweil Computer Products, 264 Third St., Cambridge, Mass. 02142.

Technical linguistic vocabulary has been avoided in this article as much as possible in order to make the algorithms accessible to a wide range of readers, including users of computers who would like to add speech output to their systems. It was impossible to avoid occasional use of "phoneme" and "phonemic", words used here simply to reflect the fact that the device used in this work has a stock of sounds which are called up simply by giving numbers ("phoneme codes"). The use of the word "phoneme" here is less precise than that used in linguistics, but this should not lead to confusion for the present purposes. Technically, a phoneme is a group of similar sounds, all of which are considered equivalent by a speaker. The kind of synthesizer used in this work produces slightly different sounds depending on the neighboring sounds, because the transitions from one sound to the next are determined by the nature of the adjacent sounds. As a result, one "phoneme" code will cause slightly different sounds to be produced, depending on circumstances, and this is close to the linguistic definition of "phoneme".

Device description

A brief description of the Votrax model VS-6 used in this work will be useful in the discussion of the specific algorithms. Strings of 8-bit codes are sent to the device, with 6 bits specifying one of 64 sounds and 2 bits specifying one of 4 pitch levels. There is no way to specify different loudness levels. Of the 64 sounds, two are actually silences or pauses, and 22 are shorter-duration versions of vowels, so that there are about 40 distinct phonemes. Diphthongs are made by combining sounds. Although the VS-6 was designed for midwestern American speech, the fact that the typical American diphthongal vowels are produced by combinations of VS-6 pure vowels means that these pure vowels are available for use in other languages. Since the internal workings of the VS-6 are inaccessible to the user, the characteristics of the phonemes and the transitions between phonemes cannot be controlled by the user. For example, certain French vowels are not obtainable from the device.

At some places in this article detailed VS-6 coding is given. The rationale for doing this is that the VS-6 is probably the most widely-distributed and successful phonemic synthesizer presently available, and many people might be able to use these implementation details. After much thought, the manufacturer's own coding symbols were used, even though these symbols are in some cases poorly chosen. In principle, the International Phonetic Alphabet could be used, but in practice the 40 sounds of the VS-6 do not correspond in a clear-cut way to IPA symbols. For example, the VS-6 symbol "CH" is not actually the sound at the end of "such" but is rather a short "sh". The sound at the end of "such" is produced when the VS-6 is given the sequence of codes "T", "CH". Enough detail is given that it should be relatively easy to implement on other synthesizers the algorithms described here.

The standard VS-6 contains an 80-sound buffer which must be loaded before speech output begins. Presumably this is necessary because the generation of the transitions between sounds involves looking ahead at successive phoneme codes. The buffer can be completely loaded in 400 microseconds, so external buffering can make it possible to sustain continuous speech. Since likely applications in the PLATO computer-based education system (Smith & Sherwood, 1976), in which this work was done, would typically involve short sentences for giving instructions, no external buffer was used.

Stress and intonation

The treatment of stress (emphasis of a syllable within a word) and intonation (falling and rising pitch contours over multi-word phrases) is similar for all five languages. In initial experiments the Votrax pitch levels were used to express stress. (Remember that the VS-6 does not permit programmable loudness variations.) For example, unstressed-vowels in Esperanto and Spanish were given pitch 2, and stressed vowels were given (higher) pitch 3. English vowels were marked with four different accent marks to specify the four possible pitches. These procedures produced unnatural-sounding sentences. In the case of Esperanto and Spanish, the sentence while understandable lacked the overall falling intonation of human sentences. In the case of English, this flaw was coupled with a great deal of confusion and indecision on the part of the typist in attempting to assign pitches rationally. English speakers are aware of the sounds and stresses in words but find it difficult to "hear" the intonation patterns of phrases (although anomalous intonation is recognized as sounding unnatural).

Attempts to aid the English typist in the assignment of intonation patterns not only simplified the typing and improved the intonation of English sentences but also led to more natural-sounding sentences in the other languages. Instead of asking the English typist to assign pitches to vowels, the typist is now asked merely to mark stressed syllables, an easy task for an English speaker, who is well aware of which syllables are stressed. The stressed and unstressed English vowels are then produced by longer-duration and shorter-duration synthesizer variants. The longer-duration vowels are perceived as "stressed" by the listener. The locations of the stressed syllables in the phrase are also the places where pitch (intonation) is shifted. The phrase begins at pitch level 3, rises to level 4 at the beginning of the first stressed syllable, then drops from level 4 to level 2 at the end of the first stressed syllable. The pitch rises to level 3 at the beginning and drops to level 2 at the end of each successive stressed syllable (except for the last one). A phrase terminated by a period falls to level 1 at the end of the last stressed syllable (see Fig. 1). This scheme has the effect of giving prominence to stressed syllables within the sentence, but with an apparent overall fall, due to the early high pitch (level 4) and the final low pitch (level 1). Most sentences sound fairly natural when given this treatment. A detail of the final fall to level 1 is that, in English, if the stressed vowel is the last vowel of the phrase a down-glide is made *on* this vowel, whereas if there are additional (unstressed) vowels a non-gliding step down is made just *after* the last stressed vowel. If there is only one stressed syllable in the sentence, a 2-3-1 contour is used.

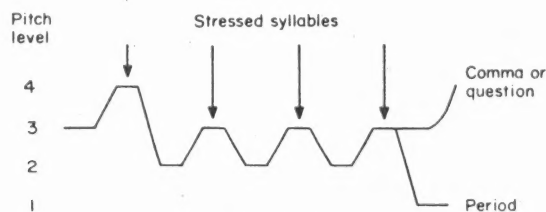


FIG. 1. A simple intonation algorithm. Stressed syllables are made prominent by having higher pitch than neighboring syllables. An overall feeling of down-drift results from the early level 4 pitch and the final level 1 pitch.

A phrase ending in a comma or question mark rises to level 4 at the very end of the phrase. However, English questions beginning with "wh" are treated as though the question mark were a period, because questions starting with such words as "what", "why" and "where" have falling intonation. These are questions which do not take a yes or no answer: normally only yes/no questions have rising intonation. To hear this, compare "Did you see it?" with "What was it?". This is also true in some other European languages, but which questions are yes/no questions is not always so conveniently marked by the spelling as it is in English. Of the other languages treated here, only Esperanto makes it easy: it has a special question word at the beginning of yes/no questions. For the other languages it would be necessary (among other things) to search a list of question words (what, why, where, etc., in the respective language).

It is occasionally advantageous to insert commas to break up long sentences into shorter phrases. The comma not only produces rising pitch but also produces a short pause. In order to give the typist flexible control over the intonation and phrasing, a semicolon can be used to specify a short pause without affecting the intonation, and a colon can be used to specify a short pause with a down-going pitch contour. Periods and question marks generate long pauses in addition to affecting the intonation pattern.

Before arriving at this intonation algorithm, two other schemes were tested. The first was patterned after Armstrong & Ward (1931), who described English (not necessarily American) sentence intonation as generally falling, but with the pitch staying constant from the start of a stressed syllable until the down-shift at the beginning of the next stressed syllable. This is in contrast to the scheme adopted here, in which each stressed syllable within the sentence begins with a rise and ends with a fall. Many listeners complained that the step-wise fall of levels of Armstrong & Ward sounded unnatural. The second scheme tested was that of Pike (1945), who characterized emotionally-neutral American intonation as a succession of 2-3-2 pitch contours around stressed syllables, with a 2-3-1 contour for the final stressed syllable. This scheme also sounded unnatural, apparently because it lacked the feeling of an overall falling pitch from beginning to end of the sentence. Finally the two schemes were combined by raising the first stress contour to 3-4-2, which gives the illusion of an overall down-drift in the pitch. In fairness to Pike, it should be pointed out that he focussed on what was emotionally distinctive about various intonation patterns, and he may have considered the normal overall down-drift as a given, without distinctive significance. Additional support for the intonation scheme adopted here may be found in Olive (1975) and Mezzalana, Rusconi & Torasso (1976).

An attempt has been made to reproduce the emotionally-neutral intonation pattern of unemphatic informational speech. It is desirable to have some way to mark words in the sentence for special emphasis in order to escape from the standard pattern. Words written in all-capitals are given special emphasis, as in "This MAN did it". The VS-6 has limited capabilities for emphasis: the present algorithm simply gives pitch level 4 instead of 3 to the stressed syllable (and, in the case of English, make a down-glide on the emphatic vowel, even if normally there would be a down-step following the vowel).

The Spanish intonation algorithm is slightly different. Instead of dropping the pitch at the end of a stressed syllable, the pitch is left high until the end of the word. The pitch then falls to level 2 and rises again at the next stressed syllable. This scheme better represents the pattern of Spanish intonation (Delattre, Olsen & Poenack, 1962). While it is common for the final stressed syllable to have low pitch, intelligibility was improved

by making the final stressed syllable high, as in English, even though this makes the last word mildly emphatic (this is also true for Italian).

Stressed vowels are given longer durations which, together with changes in pitch, make these syllables prominent. One-syllable words are not given stress in the Esperanto, Spanish and Italian algorithms, since these are usually simple function words (in particular, articles or prepositions) which are not stressed. But a one-syllable word at the end of a sentence (or at other punctuation marks) is stressed, since a simple function word cannot occur in this position. Occasionally a one-syllable word must be stressed in the middle of a sentence. With these algorithms the typist must enter such a word in all-capitals in order to give special emphasis. (In Spanish and Italian, the algorithms will stress a one-syllable word if the vowel has an accent mark.)

The last vowel in a phrase is lengthened (Umeda, 1975; Delattre, 1966), even if it is not stressed. This improves the naturalness of the speech.

Note that this simplistic treatment ignores the actual meaning of the words and phrases. No parsing is done to "understand" what is being said, and the algorithms know nothing about meaningful phrases if they are not set off by punctuation marks. Nevertheless, these simple algorithms produce natural-sounding intonation in most cases, and the typist can easily use additional punctuation marks to improve phrasing if necessary.

The Votrax model VS-6 does not change pitch at the beginning of the sound but about half a sound later. It is therefore necessary with this synthesizer to change the pitch on the phoneme code just preceding the point at which the pitch is to change, unless very long-duration vowel sounds are used. In English, at least two VS-6 codes are used for stressed vowels, with the last code representing a fairly long-duration sound. This makes it possible to produce a down-gliding pitch change on the last stressed syllable of a sentence (which is done in the absence of following unstressed syllables) simply by attaching pitch level 1 to this last code. A side-effect of this procedure is that the VS-6 speed control setting must be set rather fast to compensate for the long length of the vowels.

Witten & Madams (1977) perform some equalization of the time between stresses in synthetic English in order to produce the characteristic "stress-timed" rhythm of English. This was not attempted here, because there is no control over the duration of VS-6 consonants and only limited control over the duration of vowels. For recent careful measurements on the rhythm of (British) English, see Hill, Witten & Jassem (1978) and Hill, Jassem & Witten (1978).

Having described the treatment of stress and intonation common to all the languages, it is now possible to describe the detailed phonemic and timing characteristics of the different algorithms used for the five languages.

Esperanto

The easiest language to handle is Esperanto,[†] whose sensible design includes a spelling scheme in which each letter has only one sound (and each sound has only one letter). Moreover, stress in a word lies on the next-to-the-last vowel, without exception.

[†] Esperanto is a language designed to be easy to learn and intended for communications between people of differing native languages. See "*Teach Yourself Esperanto*", J. Cresswell & J. Hartley, Hodder and Stoughton Ltd., London (1968). (Published in the United States by David McKay Company Inc., New York.)

Written Esperanto is therefore "doubly representative" of the spoken language (both sounds and stress). A one-pass table-lookup translation is made from standard Esperanto text to synthesizer phoneme codes. Table entries corresponding to input characters include not only synthesizer codes but also numbers corresponding to special subroutines invoked by certain input characters. Vowels trigger a routine which keeps track of vowel positions. When a punctuation mark or space is encountered, the location of the next-to-the-last vowel is added to an array of stressed-vowel positions (but one-syllable words are not stressed), and the synthesizer code corresponding to the stressed-vowel position is changed to a longer-duration variant. A punctuation mark also triggers the intonation subroutine, which appends 2-bit pitch specifications to the output buffer of 6-bit synthesizer phoneme codes, using the intonation rules described earlier. The letters c, g, h, j, s, and u invoke routines which check for an attached circumflex (or hachek, an inverted circumflex, in the case of u). The supersigned letters, which are considered separate letters in the Esperanto alphabet but which involve extra character-codes in TUTOR (Sherwood, 1977) character strings, have the sounds of ch in chin, g in gem, ch in the Scottish word loch, j in the French word jour, sh in she, and w in we (but the supersigned u appears only following a or e, to form a diphthong).

The Esperanto vowels a, e, i, o, u are the same as those of Spanish and are rendered by the VS-6 phonemes AH2, AH2; EH1; E1; O1; and U. These VS-6 sounds are produced with approximately equal time durations, which is important for correct Esperanto rhythm. Stressed vowels are given longer duration (AH1, AH2; EH; E; O; and U). The VS-6 sound U is actually a little too long for unstressed Esperanto vowels, but the shorter U1 is too short. Final unstressed vowels are given the same codes as stressed vowels for naturalness (but a final "a" is not lengthened in this way because it is a little too long, and there isn't an appropriate VS-6 length for this case). If the final vowel is stressed (as in a final one-syllable word), the vowel is made even longer (AH1, AH1; EH1, EH1; E1, E; O1, O; U1, U), and the pitch changes on the second of these pairs of codes. The complete translation table is shown in Fig. 2. Note that some of the consonants require two VS-6 phoneme codes. For example, the Esperanto letter "c" is pronounced "ts", and the sequence "T", "S" must be sent to the synthesizer.

Esperanto			
a AH2, AH2	g G	k K	s S
b B	ĝ Ĝ, J	l L	ŝ SH
c T, S	h H	m M	t T
ĉ T, CH	ĥ K, H	n N	u U
d D	i E1	o O1	ŭ W
e EH1	j Y	p P	v V
f F	ĵ ZH	r R	z Z

Sample: Bonan Tagon. Mi povas paroli Esperanton.

FIG. 2. The VS-6 sounds corresponding to the letters of the Esperanto alphabet. The sample sentence says, "Good day. I can speak Esperanto." See text for details on the lengthening of stressed and final vowels.

As a specific example of how the algorithm works, consider the phrase, "la libro estas sur la tablo" (the book is on the table). The table entry for "l" is VS-6 "L", which is put in the output buffer (a list of sounds to send to the synthesizer). Associated with "a" is a table entry specifying special vowel treatment: not only are the sounds "AH2, AH2"

added to the output buffer, but the location in this buffer (sound number 2) is noted for possible stress. The space at the end of "la" is encountered next, and since only one vowel was found in the word, no stress is implied. At the end of the word "libro", however, two vowels have been seen, and the location of the next-to-the-last vowel (the "i", which is sound number 5 in the output buffer, after L, AH2, AH2, L) is added to a list of stressed syllables which will govern the intonation. Also, the short-duration "i" sound (Votrax E1) is changed to the longer-duration stressed "i" (Votrax E).

Processing continues in this way, with sounds added to the output buffer and stress locations added to a list of stressed syllables. Upon reaching the end of the phrase, the stress list shows three stress positions: the "i" of "libro", the "e" of "estas" and the "a" of "tablo". When a final period is encountered, the intonation subroutine is invoked to add pitch levels to the output phoneme codes. As described earlier, pitch levels are changed at the locations of the stressed syllables. (The pitch is actually changed on the consonant preceding the vowel, to compensate for the VS-6 delay in changing pitch. If, as is rarely the case, the stressed vowel is preceded by a vowel, it is necessary to make the pitch change on the stressed vowel itself.) The complete output buffer, which now includes both sounds and pitches, is transmitted to the synthesizer. The entire operation takes only a fraction of a second, so the sentence can be heard as soon as it is typed. Other languages described in this article are handled in similar ways, although of course there are differences in how the sounds are chosen.

Despite the extreme simplicity of the algorithm, the resulting speech output is highly intelligible. Informal tests with American and Japanese speakers of Esperanto have indicated that standard text typed directly from Esperanto books or magazines is usually understandable at first hearing. This is apparently a property not merely of the phonetic spelling but of the simplicity of the sounds.

Spanish

Spanish spelling, like that of Esperanto, is "doubly representative" of the language, both of sounds and of stress. Simple rules determine the position of the stressed vowel, and when these rules are violated it is *obligatory* to place an acute accent mark on the irregularly-stressed vowel. For example, the Spanish word "capitán" (captain) has an accent mark on the third vowel to prevent the normal stress rules from stressing the second syllable. There is one exception to this practice: capitalized vowels do not receive accent marks, so the spelling of "Los Angeles" does not show that the "A" is the stressed vowel. In order for the Spanish text-to-speech algorithm to produce the correct intonation pattern the typist must place an accent mark on capital letters where necessary, which violates normal spelling practice. The only other common irregularity is that Mexicans spell the name of their country "México", which by standard Spanish spelling rules would be pronounced "Méksiko" but which is actually pronounced "Méhiko". However, people in other Spanish-speaking countries spell the country "Méjico", in agreement with the pronunciation, and this is the form that should be used for producing correct synthetic speech.

There are minor differences in Spanish pronunciation from country to country in Latin America, and the Spanish of Spain has a "th" sound where American Spanish pronounces "s" or "z". For definiteness, an algorithm for Mexican Spanish is described here. This form is easily understood by other speakers of Spanish.

The phoneme translation for Spanish is more complicated than that for Esperanto, in that one letter can have several sounds, depending on context. For example, the letter *c* (which *always* has the sound "ts" in Esperanto) is associated with three different sounds in Spanish: "ch" in "Chile", "s" in "centavo", and "k" in "Caracas". The letter *following* the *c* determines the sound. A following *h* causes a "ch" sound, a following *e* or *i* causes an "s" to be produced, and any other following letter makes the *c* have the "k" sound. Similar contextual rules are all completely regular, without exceptions. Detailed rules are given in Appendix A. Among the major European languages, Spanish is unique in having a spelling which is almost as regular as that of the constructed language Esperanto, including stress.

A table-lookup scheme similar to the Esperanto algorithm is used. The Spanish vowels are the same as those of Esperanto. The vowel-handling subroutine in addition to keeping track of vowel positions also checks for an associated accent mark, which will override the normal stress assignment. When a space or punctuation mark is encountered, stress rules are used to determine the stressed vowel (if there was no accent mark). These stress rules state that the next-to-the-last vowel is stressed if the word ends in a vowel, "n", or "s"; otherwise the last vowel is stressed. One-syllable words with no accent mark are not stressed by this algorithm except at the end of a phrase, since these are almost always simple function words such as prepositions or articles.

As with Esperanto, the resulting speech output is highly intelligible. Informal tests with several native speakers of Spanish have indicated that standard text typed directly from Spanish books or magazines is understandable at first hearing (though an occasional unaccented capitalized vowel will not be stressed properly). This synthetic Spanish has a North American accent due to the particular sounds in the VS-6. In particular, the *r* lacks the tap or trill of the Spanish *r* or double *r*, and the *v* is not quite the same as the Spanish *v*. However, these defects do not significantly hinder intelligibility.

Italian

Italian spelling is not "doubly representative" of the sounds, because stress is only occasionally indicated by an accent mark, and in the absence of an accent mark there is no simple rule for determining the stressed syllable. Moreover, *e* and *o* when stressed have two possible pronunciations which can change meaning. For example, "pésca" ("fishing", with the *é* pronounced approximately like the vowel in "say") and "pèsca" ("peach", with the *è* pronounced as in "set") have the same spelling, without accent marks, except in dictionaries, where the two pronunciations are distinguished by the two kinds of accent marks. The word "pòsta" (feminine form of the adjective meaning "placed", with the *ó* pronounced as in "toe") and the word "pòsta" ("mail", with the *ò* pronounced as in "awe") also have accent marks only in dictionaries. The unstressed Italian vowels *a*, *e*, *i*, *o*, *u* are rendered by VS-6 codes AH2, AH2; A; E1; O1; U. The longer-duration stressed vowels are given by AH1, AH2; A; E; O; U. The variants *è* and *ò*, which occur only in stressed syllables, are given by VS-6 codes EH and AW1.

The treatment of stress is similar to that of the Spanish algorithm: the next-to-the-last vowel is stressed unless some vowel has an accent mark. One-syllable words are not stressed except at the end of a phrase. If the next-to-the-last vowel is not the stress location (or if it is pronounced as *è* or *ò*), the typist must place an appropriate accent on

the stressed vowel, even though standard Italian texts would not show an accent mark in the word. This means that text taken directly from Italian books or magazines will be stressed incorrectly and have slightly incorrect vowel sounds in some words.

As in Spanish, the pronunciation of Italian letters depends on context. Details are given in Appendix B.

The resulting output is highly intelligible. Unlike Esperanto and Spanish, Italian text from books or magazines requires additional accent marks. Informal tests with a native speaker of Italian indicate that the synthetic Italian is understandable at first hearing.

Russian

While many unaccented Italian words have stress on the next-to-the-last vowel, no simple stress rule covers the bulk of Russian words. Also, many one-syllable Russian words are important words which must be stressed for the intonation algorithm, whereas one-syllable words in Esperanto, Spanish and Italian are usually simple function words which are not stressed. Stress location in a Russian word affects vowel sounds: for example, the Russian letter *o* is pronounced three different ways, depending on whether it is stressed, immediately precedes the stressed syllable, or appears in some other location. It is therefore necessary for the typist to place an accent mark on the stressed vowel in each word. This practice is also followed in elementary Russian-language readers for foreigners. Unlike the practice in such readers, two different accent marks are needed, where the second can be used to determine vowel sounds without influencing the intonation algorithm. This is necessary because some function words, such as prepositions, can have several syllables but should not be stressed. The importance of proper stress, coupled with the effect of stress position on vowel sounds, means that the Russian algorithm presented here will *not* handle well text taken directly from Russian books or magazines, where stress is not marked.

The sounds of Russian consonants are accurately conveyed by the normal spelling except in one minor but common situation. There are grammatical word endings which are spelled -*ogo* and -*ego* (in Latin transliteration) but which are pronounced -*ovo* and -*evo*. A scan of a large dictionary indicates that final -*ogo* and -*ego* should always be pronounced with a "v" except in the words "ogo" (O Ho!), "mnogo" (much), "dorogo" (dear), "strogo" (strictly), in prefixed words ending in -*mnogo*, -*dorogo* or -*strogo*, and in the scientific term "Mogo" (Moho, the region between the earth's mantle and crust). In positions within a word, -*ogo*- and -*ego*- are pronounced with a *g* except in words ending with "-egosia" (reflexive past participles) or beginning with "sego-": "segodnia" (today), "segodniashnii" (today's) and "segoletka" (this year's brood).

There are rather complex contextual rules for translating from Russian text to speech. The details are given in Appendix C.

Several teachers of Russian and several native Russian speakers have heard the synthetic Russian and have found it to be very intelligible. Since the Votrax VS-6 was designed for midwestern American speech, and since the Russian sound system is quite different and complex, it is perhaps surprising that the synthetic Russian is understandable. Like Italian, the Russian algorithm requires modification of standard text in order to work well.

English

As was mentioned earlier, it is possible to produce high quality English from standard text, but this requires a large dictionary of words and roots plus some parsing of the sentence, and the translation by a powerful computer takes many seconds to produce one second of speech. Not only does English spelling not indicate sounds unambiguously (compare "break" and "bleak"), it also does not indicate stress (compare the noun *próject* with the verb *próject*). Where Esperanto and Spanish have doubly representative spellings (in both sounds and stress), Italian and Russian spellings are singly representative (sounds only), and English spelling is doubly unrepresentative. A *simple* algorithm to produce good-quality speech from English text is possible only if phonetic text is used, with stress marked in the words. The question remains whether a phonetic spelling scheme would be useable by people who are not trained in phonetics.

World English Spelling (WES)

a	fat	o	on (between aa and au)
aa	faather (father)	oe	toe
ae	Mae West	oi	oil
ar	far	oo	too
au	taut	or	for
b	but	ou	out
ch	chum	p	pet
d	dig	r	run
e	set	s	set
ee	see	sh	shed
er	gather	t	tin
f	fat	th	this
g	gum	thh	thhing (thing)
h	hat	u	up
i	in	ue	hue
ie	tie	ur	fur (a long "er")
j	jam	uu	buuk (book)
k	kit	v	van
l	let	w	win
m	met	wh	when
n	net	y	yes
ng	sing	z	zoo
nk	sink	zh	vizhun (vision)

Separate ambiguous combinations with a period, as in "mis.hap" (to prevent the "sh" from having its usual sound). Capitalize the beginning of stressed syllables, as in "Forskör and Seven Yirz uGoe, our Faatherz Braut Forthh on this Kontinent u Noo Naeshn, kunSeevd in Liberty, and Dedikaeted too thu propuZishun that Aul Men are kreeAeted Eekwul."

FIG. 3. The World English Spelling (WES) scheme uses the Roman alphabet to spell the sounds of English. Note that the "long" vowels (ae, ee, ie, oe, ue) end in e, which is reminiscent of "silent e" in normal spelling.

A study of the extensive literature on English spelling reform uncovered a suitable alphabet, the Word English Spelling (WES) alphabet (Dewey, 1971) (see Fig. 3). This alphabet was developed by the American branch of the spelling reform movement. Other alphabets are reviewed in Haas (1969). It has been found experimentally that most people quickly become proficient with WES after typing and hearing WES

sentences, and that they find it a pleasant and reasonably natural scheme (this is in a situation where they have the immediate gratification of hearing the results of their typing). Moreover, WES text is quite readable, which is an important feature of text appearing in a computer program. The next paragraph is written in WES, with the beginnings of stressed syllables capitalized, an addition to WES necessary for informing the algorithm as to which syllables should be stressed. Accent marks could be used instead, but one of the main advantages of WES is the fact that one can use a standard typewriter, which may not have accent marks. Not all stressed syllables are marked in the following paragraph: the sentence sounds more natural if some relatively unimportant words do not have any stress marked. Notice that one organizing principle of WES is that the "long" vowels are spelled *ae*, *ee*, *ie*, *oe*, and *ue*, by analogy with the final "silent e" so characteristic of English spelling. Refer to Fig. 3 for help in reading the following paragraph.

thu Difrent Kiendz uv foeNetik Alfubets split intoo
 Too Maen Groops: thoez uezing thu Standard Roemun Leterz
 (such as This Wun), and thoez uezing kumPleetly Noo Leterz,
 az in thi internashunl foeNetik Alfubet, or in thi inIshul
 Teeching Alfubet. thi adVantej uv Ueizing thu Standerd Roemun
 Leterz iz that Tieprieterz and kumPueter Sistemz kan Handl
 them Eezily. uMung such Alfubets, Dublyoo Ee Es haz thu
 Ferther adVantej uv beeing perTikuelerly Eezy too Reed and
 Riet.

Since there are only 26 Roman letters, two-letter combinations (and one three-letter combination, "thh") are used along with single letters to express the approximately 46 sounds of English. (One could get by with as few as 38 symbols, in that the WES symbols, *ar*, *ie*, *nk*, *oi*, *or*, *ou*, *ue* and *ur* could be built up out of other forms. For example, *ar* is equivalent to *aar*, *ue* is essentially equivalent to *yoo*, *nk* is really *ngk*, etc.) Occasionally a misleading association of neighboring letters must be broken up by inserting a period: for example, the word "mishap" must be spelled "mis.hap" to prevent the "sh" from having its customary sound. This happens rarely, and the preceding WES paragraph did not contain any example of this.

For more than a hundred years many Roman-letter phonetic alphabets have been proposed for English, but none is so readable as WES nor so free of letter-pairing problems, while at the same time giving an accurate representation of the sounds. Some of these other alphabets achieve more natural-appearing text than WES, but at the cost of introducing special, non-phonetic spellings for 20 or 30 common words. For example, some such alphabets permit "of" to be spelled "of" instead of "uv". The slight gains for the reader seem outweighed by the uncertainty induced in the writer as to when a non-phonetic spelling is permitted. One major scheme, the "Regularized English" of Wijk (1959), attempts to preserve natural appearance at the high cost of very complex rules, including some outright irregularities (as in the pronunciation of *g* in "get" and "gem"). Many of the alphabets use *dh* and *th* for the WES *th* and *thh*. While there is logic on the side of using *dh* for the voiced form, its use makes a large number of common words look strange: "this" and "that" become "dhis" and "dhat". The unusual three-letter WES combination *thh* is less noticeable to the reader, partly because it is not nearly as common in English text ("bathh", "thhink", etc.). Some

alphabets avoid the frequent appearance of words such as "dhis" and "dhat" by including "this" and "that" in the list of non-phonetically-spelled irregular words.

Except for such relatively minor details, most of the proposed Roman-letter alphabets are quite similar to each other, because they all attempt to make text look as similar to traditional spelling as possible. For example, all such alphabets have the letter "a" stand for the sound in "fat", because this is by far the most common sound of the letter "a" in traditional spelling, at least in stressed syllables (its pronunciation in unstressed syllables is highly irregular). An example of where these alphabets differ is in their use of the symbols "oo" and "uu". Some alphabets, including WES, assign to "oo" the sound of "too" and to "uu" the sound of "put". Others reverse these assignments, associating "oo" with the sound in "book" (or "put"), which leaves "uu" to stand for the sound in "too". The problem here is that "oo" in the traditional spelling is used about as often for the sound of "book" as it is for the sound of "too", so that the appropriate assignment is not so obvious as is the case for the letter "a". Ultimately a choice among such alphabets is a matter of taste and judgement, but WES seems the most appropriate alphabet for the particular purpose of speech synthesis.

Unstressed vowels are difficult to treat properly in English. A baby cat is sometimes called (in WES spelling) a "kitn", "kiten", "kitin" or "kitun". The WES paragraph shown earlier might appear incorrect to some speakers, who would make other choices for the WES spellings of their own pronunciations. (Another way of saying this is to point out that WES does provide a fair amount of latitude for individual or regional pronunciation differences.) In the spelling reform literature insufficient attention has been paid to the problems of unstressed vowels, and reformers have often tacitly assumed that the native speaker will "know" how to pronounce the unstressed vowels, even if spelled "incorrectly". Such additional knowledge is lacking to an electronic voice synthesizer, so that such spellings as "about" for "ubout" or "together" for "together" (or "tugether" or "tigether"), often seen in the spelling reform literature, are unacceptable in speech synthesis.

A few "synonymous" spellings have proved useful in making it easier for the typist to create WES sentences. The letter c appears only in the combination ch in WES, but it is helpful to the typist to permit c when followed by e, i, or y to stand for the sound "s", and everywhere else to stand for the sound "k" (except before h, of course). Thus the words "cent", "cinder", "fancy" and "cat" retain their normal spellings in this modified WES. Since there is no ambiguity in these synonymous spellings, what would otherwise be an "erroneous" WES spelling is treated appropriately, at low processing cost. Other helpful synonyms are ai and ei for ae (fail, rein), aw for au (law), dg for j (budget), ea for ee (each), ow for ou (cow), oy for oi (boy), qu for kw (quiz), and x for ks (next). These additions are at variance with one of the main tenets of WES, that of having only one spelling for each sound. However, in actual use it is very helpful to the inexpert typist that these common alternative forms are accepted by the speech algorithm. For synthesizing speech all that is required is consistency: there is no compelling reason to rule out additional spellings, as long as these are unambiguous. It should be added that these additional symbols are about the only ones outside WES which are relatively unambiguous in their sounds. While it is true that "x" is sometimes pronounced "gz" (as in "exaggerate"), that "ow" is sometimes pronounced "oe" (as in "low"), and so on, the assigned sounds are those which in the great majority of the cases are associated with these spellings in English, whereas other candidates for inclusion turn out to be quite

irregular in their sounds. Wijk (1959) includes a very useful survey of the various spellings of English sounds and of the various sounds of English spellings.

While the WES ae, oe, and oo represent slightly diphthongized vowels, it is useful to let ay, oa and eu stand for the un-diphthongized parts of these vowels, permitting slight phonetic distinctions between "sae" and "say", between "coet" and "coat", and between "boot" and "beut". These distinctions can be used for special effects, such as synthesizing foreign words. Another use is in making very long diphthongs. For example, the WES ae is a diphthong, with some ee at the end. If a long version of ae is desired, it can be typed as ayayayee. Typing aeaeae wouldn't work, because the ee sound would intrude several times, as in ayeeyeeayee.

There is a minor problem with words such as "there" and "air". The WES spellings are "the.r" and "e.r", which look clumsy (the period is needed to prevent the er from having the sound it has in "gather"). An appropriate extension or modification of WES is to write these words as "thaer" and "aer" (or "thair" and "air", or "their" and "eir"). The "long ae" hardly ever occurs before "r" in English, so there is no real problem in introducing this modification into the standard WES scheme. The new rule is simply that a following r converts ae (including its synonyms ai and ei and its near synonym ay) to the WES sound of e (as in set). There is a similar problem with words such as "ear" and "here", which people tend to spell in WES as "eer" and "heer", although the correct phonetic spellings are "ir" and "hir". Since the "ee" sound hardly ever occurs before "r" in English, there is no harm in helping the typist by interpreting "eer" or "ear" as meaning "ir".

While WES is easy for users, it could also be a helpful "documentation" alphabet among people who are working on English speech synthesis, who are often handicapped by not having the International Phonetic Alphabet available on their typewriters or computer systems. Roman-like alphabets used for this purpose by researchers have often been quite unnecessarily hard to read and write, and WES would be a big improvement for documenting English speech algorithms. For an example of a highly unreadable alphabet, see the ARPABET alphabet reviewed in Rice (1976).

It should also be mentioned that many Votrax users have hand-coded individual words by painstakingly choosing specific Votrax coding variants. While in principle WES spelling cannot produce speech as good as could be produced by careful hand-tuning of every word, it seems to turn out in practice that WES results are often indistinguishable from the results of typical hand-coding methods. Not only is typing a WES sentence a much more efficient way to enter text, it also assures some overall consistency from word to word in the way sounds are chosen. This probably results in better intelligibility, since the machine has a consistent "accent". Also, storing the text in WES format can make it feasible to handle new devices which have different coding schemes, or to handle several different devices on the same computer system. The use of capitalization for marking stress, together with the intonation algorithm described here, makes it feasible to store individual WES words in a data base and assemble natural-sounding sentences by program, simply by concatenating these words.

Figure 4 shows the VS-6 codes for WES sounds, including the codes for long-duration and extra-long-duration vowels. Vowels are lengthened if stressed or when followed by one of the voiced consonants b, d, g, j, v, or z. The reason for lengthening vowels when a voiced consonant follows is that this helps distinguish "bad" from "bat", "slab" from "slap", etc. If the vowel is specially emphasized (by the use of all-capitals,

		English	
a	AE1 (AE1, AE1; AE1, AE)	o	AW2, UH1 (AW1, UH1; AW, UH1)
aa	AH1 (AH2, AH1; AH1, AH1)	oe	O2, U1 (O1, U; O, U)
ae	A2, E1 (A1, E1; A, E1)	oi	O2, EH3, E1 (O1, EH3, E1; O, EH3, E1)
ar	AH1, R (AH, R; AH, R)	oo	IU, U1 (IU, U; IU, U1, U)
au	AW1 (AW1, AW1; AW1, AW)	or	O2, R (O1, R; O, R)
b	B	ou	AH2, O1 (AH1, O1; AH, O1)
ch	T, CH	p	P
d	D	r	R
e	EH1 (EH2, EH1; EH2, EH)	s	S
ee	E1 (E1, E1; E1, E)	sh	SH
er	ER (ER, R; ER, ER)	t	T
f	F	th	THV
g	G	thh	TH
h	H	u	UH2 (UH2, UH1; UH2, UH)
i	I1 (I2, I1; I2, I)	ue	Y, U1 (Y, U; Y, U1, U)
ie	AH2, I1 (AH1, I1; AH, I1)	ur	ER, ER (ER, ER; ER, ER)
j	D, J	uu	OO1 (OO1, OO1; OO1, OO)
k	K	v	V
l	L	w	W
m	M	wh	H, W
n	N	y	Y
ng	NG	z	Z
nk	NG, K	zh	ZH

FIG. 4. VS-6 codes for WES sounds. Normal, long, and extra-long versions of vowels are shown. The long version is used if the vowel is stressed or followed by one of the voiced consonants b, d, g, j, v, or z. If the vowel is specially emphasized (all-capital letters), the extra-long version is used and the first sound is doubled.

as in WOW), the extra-long-duration variant is chosen, with the first listed sound doubled by being output twice.

Comparisons and speculations

All major human languages are adequate languages in their corresponding societies, are learnable by the young, and most of them have rich written and oral literatures. However, languages may differ in the ease with which certain technological operations can be performed, such as in the manipulations involved in going from text to speech or from speech to text, using simple machines. The present work gives one objective basis for rating several major languages in these terms.

There is a spectrum of difficulty and complexity of the text-to-speech algorithms for these five languages, running from the very simple Esperanto to the quite complex English. Accompanying the increasing complexity is the necessity of making increasing modifications to the standard spelling of these languages. If one considers what is necessary to handle standard written English rather than the stress-marked phonetic English considered here, it is clear that the English sound system and the English writing system are much more difficult for simple phonemic speech synthesis than are those of Esperanto or Spanish.

Increasing complexity seems to be coupled with decreasing ease of understanding. The synthetic Esperanto is very easy to understand (even though the translation from

standard Esperanto text to synthesizer codes is extremely simple), while understanding the synthetic English requires some effort (even though many tricks were pulled in the algorithm to try to cope with the difficulties). To make further improvements in the synthetic English may require more than merely improving the individual sounds produced by the synthesizer. It may be that the underlying structure of spoken English simply does not correspond well to the structure of a simple phonemic synthesizer, whereas the structure of Esperanto matches very well. Subtle changes in relative timing, pitch, and loudness of English syllables seem very important to intelligibility, whereas it is observed experimentally that these aspects matter hardly at all in the intelligibility of Esperanto, Spanish and Italian. One might picture a phonemic synthesizer as a kind of typewriter of sounds, with each sound appearing neatly "printed" in the air. This neat "printing" is similar to the distinct syllable-by-syllable pronunciations of Esperanto, Spanish and Italian, which are therefore easy to "read" when "printed" this way. But the sound structure of spoken English in this analogy is not at all like printing—it is rather like the complex script of shorthand or of beautiful Persian handwriting, which are difficult to produce on a typewriter. It is hard to imagine a *simple* device imitating the complex patterns of English, just as it is difficult to imagine a simple typewriter for Chinese. There do exist typewriters for Chinese, and there do exist high-quality English speech synthesis systems, but they are not simple devices.

It seems likely that similar considerations hold true in the other direction, that of speech-to-text, where one would like to speak to a computer and see the speech printed out on paper. This is very difficult in English (Reddy, Erman, Fennell & Neely, 1976), where the acoustic information must be supplemented by grammatical and contextual information in order to recognize sentences, even in very limited contexts. Identification of isolated words in a small vocabulary can be done with rather high accuracy, but the conversion of general connected English speech to text is an unsolved problem. Perhaps speech-to-text would be simpler in a language such as Esperanto, both because the sound system is simpler and because the spelling is regular.

It should be easier for a simple device to identify Esperanto vowels than to identify English vowels. The WES vowels ee, i, e, a, aa, au, oa, uu, oo, u and er are quite close to each other in their acoustical properties, whereas Esperanto and Spanish lack WES i, a, au, uu, u and er, leaving the remaining ee, e, aa, oa and oo widely separated and therefore more easily distinguished. The Esperanto and Spanish vowels are never "reduced" to obscurity in unstressed syllables, as they are in English. Also, in Esperanto the regular stress and grammatical word endings would help signal the boundaries between words. Spanish is not quite as good in these respects, since Spanish sometimes has several ways to spell the same sound ("ciento" or "siento", "gallo" or "gayo", at least in Mexican pronunciations, and silent "h" is common), Spanish has fewer, less regular grammatical word endings, and it can have syllable stress anywhere in a word. While Italian text-to-speech has ambiguities, Italian speech-to-text rules are so regular that the Italian language does not have a distinct verb "to spell" (there is a silent "h", but it occurs only in the verb "to have", which should cause little difficulty). Moreover, almost all Italian words end in a vowel, which should help in identifying word boundaries. In Russian, as in English, unstressed vowels tend to be reduced in such a way as to be ambiguous in their spelling and unclear in their sounds. Also, Russian voiced consonants can be devoiced by a following unvoiced consonant (and vice versa). This provides another source of possible ambiguity. It is interesting that there is an

asymmetry in each of these natural languages in the regularity of text-to-speech and speech-to-text operations (except that English is symmetrically highly irregular in both directions!). Esperanto is completely regular in both directions.

Other languages

Some languages can not be handled by the VS-6 synthesizer due to a lack of certain sounds. For example, some French and German vowels are missing (but a German version of the Votrax is available). One could imagine having a much larger stock of sounds, but a better scheme would be to make the synthesizer programmable. One could then load the parameters for a particular language or dialect before sending strings of phoneme codes to the synthesizer. The simple algorithms described here could be incorporated into the synthesizer apparatus, so that actual text rather than phoneme codes could be sent to the device.

There are undoubtedly other languages for which a simple algorithm could produce relatively high-quality speech. The languages treated in this work were chosen due to personal acquaintance. In this laboratory, Hugo Feugen has written an algorithm which translates standard German text into a phonetic (IPA) representation, which is then used to drive the VS-6 (limited by the lack of certain sounds). This German algorithm is much more sophisticated than the procedures described in this paper. Parts of speech, prefixes, suffixes, and roots are identified in order to assign correct sounds, stress, intonation, and rhythm. Complex characteristics are assigned to each German letter, and a large number of rules are used to add, delete, or change these characteristics according to context.

Mandarin Chinese has been synthesized on a VS-6 (Suen, 1976). *

Applications

There are many situations in computer-based education where it would be useful to add synthetic speech to the presentation of lesson material. Applications include speaking to a young child about words and pictures shown on the child's screen, teaching the blind, giving English or foreign-language dictation drills, or describing a complex physical phenomenon while simultaneously simulating the phenomenon on the science student's screen.

WES, Esperanto, IPA and Spanish algorithms have been built into the PLATO system in such a way that authors can easily add speech to their computer-based educational materials. The TUTOR language now has a "say" command for producing speech from text (a "saylang" command is used to specify WES, Esperanto, IPA or Spanish). An option in the PLATO program editor permits the interactive composing of "say" statements. Donald Shirer has added mechanisms to the WES algorithm to interpret mathematical expressions, which greatly enhances the capability of the "say" command. Hugo Feugen designed and Donald Shirer implemented the IPA option, which gives an author complete control over synthesizer sounds, pitch, and duration (as with Russian, a special character set is used). Perhaps the most exciting aspect of these facilities is the ease with which character strings can be created and concatenated by program to produce algorithmically-constructed sentences. These algorithms require about one seven-hundredth of a second of processing on a CDC 6400 computer to

And later: Cheng, C.C., & Sherwood, B. (1981). Technical Aspects of Computer-Assisted Instruction in Chinese. *Studies in Language Learning* 3, 156-170.

produce one second of good-quality speech. In other words, the algorithms run at a speed of about 700 times real time speed.

I thank Paul Tenczar for interesting me in this area, Mike Williams for hardware support, Phil Cooper for help with the Russian algorithm, Joseph Allen, Jr. and Spurgeon Baldwin for help with the Spanish algorithm and useful discussions, Elaine Paden for help on intonation, and Hugo Feugen for useful discussions. I thank Control Data Corporation and the University of Illinois Aviation Research Laboratory for loans of VS-6 synthesizers.

References

- ALLEN, J. (1976). Synthesis of Speech from Unrestricted Text. *Proceedings of the Institute of Electrical and Electronic Engineers*, **64**, 433-442.
- ARMSTRONG, L. E. & WARD, I. C. (1931). *A Handbook of English Intonation*. Cambridge: W. Heffer & Sons Ltd.
- DELATTRE, P. (1966). Syllable length in English, Spanish, German and French. *International Review of Applied Linguistics*, **4**, 183-198. (Note that this deals with syllable length, not vowel length.)
- DELATTRE, O., OLSEN, C. & POENACK, E. (1962). A comparative study of declarative intonation in American English and Spanish. *Hispania*, **45**, 233-241.
- DEWEY, G. (1971). *English Spelling: Roadblock to Reading*. Columbia University, New York: Teachers College Press.
- ELOVITZ, H. S., JOHNSON, R., MCHUGH, A. & SHORE, J. E. (1976). Letter-to-sound rules for automatic translation of English Text to Phonetics. *IEEE Transaction on Acoustics, Speech, and Signal Processing*, **24**, 446-459.
- HAAS, W. (Ed.) (1966). *Alphabets for English*. Manchester: Manchester University Press.
- HILL, D. R., JASSEM, W. & WITTEN, I. H. (1978). A statistical approach to the problem of isochrony in spoken British English. *Research Report Number 78/27/6*. Department of Computer Science, The University of Calgary.
- HILL, D. R., WITTEN, I. H. & JASSEM, W. (1978). Some results from a preliminary study of British English speech rhythm. *Research Report Number 78/26/5*. Department of Computer Science, The University of Calgary.
- KUTSCH, J. A. (1976). A talking computer terminal. *Report 76-815* of the Department of Computer Science, University of Illinois, Urbana-Champaign. (The algorithm of McIlroy (1974) was implemented in a microprocessor to drive a Votrax.)
- MAKHOUL, J. (1975). *Linear prediction: a tutorial review*. *Proceedings of the Institute of Electrical and Electronics Engineers*, **63**, 561-580.
- MCILROY, M. D. (1974). Synthetic English speech by rule. *Computing Science Technical Report 14*. Bell Laboratories.
- MEZZALAMA, M., RUSCONI, E. & TORASSO, P. (1976). Automatic generation of pitch contours for speech synthesis. *IEEE Conference on Acoustics, Speech, and Signal Processing*, p. 82. (Italian pitch contours are presented.)
- OLIVE, J. P. (1975). Fundamental frequency rules for the synthesis of simple declarative English sentences. *Journal of the Acoustical Society of America*, **57**, 476-482.
- PIKE, K. L. (1945). *The Intonation of American English*. Ann Arbor, Michigan: University of Michigan Press.
- REDDY, D. R., ERMAN, L. D., FENNELL, R. D. & NEELY, R. B. (1976). The Hearsay-I speech understanding system: an example of the recognition process. *IEEE Transactions on Computers*, **25**, 422-431.
- RICE, L. (1976). Hardware and software for speech synthesis. *Dr Dobb's Journal of Computer Calisthenics & Orthodontia*, April 1976, p. 6.
- SCHAFER, R. W. & RABINER, L. R. (1975). Digital representations of speech signals. *Proceedings of the Institute of Electrical and Electronic Engineers*, **63**, 662-677.
- SHERWOOD, B. (1977). *The TUTOR Language*. Minneapolis, Minn.: Control Data Education Company. (TUTOR is the author language of the PLATO computer-based education system.)

- SMITH, S. & SHERWOOD, B. (1976). Educational uses of the PLATO computer system. *Science*, **192**, 344-352.
- SUEN, C. Y. (1976). Computer synthesis of Mandarin. *IEEE Conference on Acoustics, Speech and Signal Processing*, 698-700.
- UMEDA, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America*, **58**, 434-445.
- WIJK, A. (1969). Regularized English. *Acta Universitatis Stockholmiensis VII*. Stockholm: Almqvist and Wiskell. (This work is summarized by Wijk in Haas, 1969).
- WITTEN, I. H. & MADAMS, P. H. C. (1977). The telephone enquiry service: A man-machine system using synthetic speech. *International Journal of Man-Machine Studies*, **9**, 449-464.
- YEAGER, R. (1976). Using audio with CAI: experiences of the PLATO elementary reading project. *Proceedings of the Association for the Development of Computer-based Instructional Systems (ADCIS)*, August 1976, p. 227 (Minneapolis).

Appendix A: Details of Spanish

Both spelling rules and phonetic rules are required to translate Spanish text to (Mexican) Spanish speech. An example of a spelling rule is that "ll" is the spelling of the sound "Y" (as in "llama"). An example of a phonetic rule is that written "n" is pronounced "M" when followed by a "labial" consonant (i.e. a consonant made with the lips, including b, f, p and v, though not m): "un beso" is pronounced "umbeso" and "triunfo" is pronounced "triumfo".

A table was set up to contain in a convenient format the relevant phonetic properties of the Spanish letters. For each letter, a table entry tells whether the letter is a vowel (a, e, i, o, u), a "front" vowel (e, i), a "nasal" consonant (m, n), a "voiced" consonant (b, d, g, l, m, n, r, v), whether it causes a following "r" to be pronounced "rr" (l, n, s) and whether a word terminating in the letter is stressed on the next-to-the-last vowel rather than on the last vowel (a, e, i, o, u, n, s). Thus the table entry corresponding to the letter "e" shows the letter marked as a vowel, as a front vowel and as a letter which, when found at the end of a word, indicates that the next-to-the-last vowel is stressed. (Note that the stress rule can be overridden by an explicit accent mark, as in "capitán" and "árbol".) In addition to these phonetic properties, each table entry contains the usual VS-6 sounds for the letter and a pointer to the subroutine used to process this letter.

The letters a, e and o are processed by a vowel-handling subroutine which keeps track of the last two vowels encountered in a word. At the end of a word, if no accent mark was present in the word, the stress-related marker in the table entry for the last letter is used to assign stress to the last or to the next-to-the-last vowel. At the end of the phrase or sentence these stress locations are used to determine intonation. The letters i and u undergo some additional processing. Either of these letters if lacking an accent mark and if followed or preceded by a vowel becomes a "glide": i becomes y and u becomes w in such contexts. The combination "ui" is pronounced "wi", and the combination "iu" is pronounced "yu". One way to produce this effect is to overwrite the properties of the first vowel with those of the corresponding glide, so that when processing the second letter the preceding letter no longer appears to have been a vowel. (As a matter of programming efficiency it is convenient always to have in hand the properties of the preceding letter.) There exist verb infinitives ending in -air, -eir, and -oir with stress on the i, which prevents the i from changing into a y. Since in these cases there is no accent mark on the i, it is necessary to check for a following r followed by a space or punctuation mark signalling the end of the word. If the i or u does not turn into a glide, the algorithm branches to the normal vowel-handling subroutine. Vowel

durations for unstressed, stressed and final vowels are handled in the same way as in the Esperanto algorithm.

The letters b and v are completely equivalent and are therefore processed by the same subroutine. This routine checks whether the preceding sound was a nasal, in which case the VS-6 sound is "B"; otherwise the sound is approximated by VS-6 "V". At the beginning of a sentence, or after a pause, b or v is also pronounced as "B". A convenient way to incorporate this additional context is to simulate a nasal property for a non-existent "preceding letter" in these situations. After deciding on the pronunciation of the b or v, the algorithm invokes the "labial" subroutine, which checks whether the preceding *sound* was "N", in which case the "N" is overwritten with "M" (as in "un beso" becoming "umbeso"). The letters f and p also invoke this labial routine.

The treatment of the letter d is similar to that of b and v. A nasal, or the beginning of a sentence, or a pause make a following d have the sound "D" (unlike b and v, a preceding l will also have this effect), whereas in other situations the sound is VS-6 "THV" (voiced th).

The letter c generates VS-6 codes "T", "CH" if followed by h (as in "Chile") and "S" if followed by a front vowel ("centavo", "cita"). Otherwise the sound is "K", and in this case a preceding sound of "N" is changed to "NG" (as in "un conjeo" or in "incompleto"). The sound of "K" can also be spelled "qu", with the same effect on a preceding "N" sound (as in "un queso").

The letter g has the sound "H" if followed by a front vowel ("general", "gitano") and the sound "G" otherwise ("gaucho"), in which case a preceding "N" sound is changed to "NG", as in "un gaucho". (In the same situations where b or v is pronounced "V" instead of "B", g should have a sound not available in the VS-6, and one must make do with "G".) If "gu" is followed by a front vowel the u is silent ("guerra"), but the u is normal if it bears an umlaut ("güelfo").

The letter h is always silent. The combination "ll" is equivalent to "y". If m comes at the end of a word ("álbum") it is treated as though it were an n. If n has a tilde ("año") the resulting "ny" sound is produced with the VS-6 codes "N", "I3" (the latter being essentially a short y).

There is a contrast between a "single-tap" r and a "trilled" r, as in "caro" and "carro". (Neither of these sounds is available in the VS-6, and for lack of anything better one must use VS-6 "R" and double "R" for these sounds.) The trilled r sound is indicated not only by the "rr" spelling but also when the preceding letter is l, n, or s ("alrededor", "honra", "Israel") and at the beginning of a word ("rico"). A convenient way to force a word-initial r to change to "rr" is to merge the rr-property of l, n, and s with the other properties of the preceding sound whenever a space or punctuation mark is encountered. Thus when the space in "cinco ricos" is encountered, the properties of the preceding "o" are modified to indicate that a following "r" should be changed to "rr". Alternatively one could look ahead for an initial r whenever a space or punctuation mark is encountered.

Like b and v, the letters s and z are completely equivalent in Mexican Spanish. The sound is "S" unless there is a following voiced consonant ("desde", "los dientes"), in which case the sound is "Z". A simple way to handle this phonetic rule is to make a check, just before adding a VS-6 code to the output, as to whether the previous sound was "S", then change the "S" to "Z" if the present sound is a voiced consonant. Note that the rule must be based on sounds, not on letters. For example, in "los generales"

the letter g turns out not to be a voiced consonant but has the sound "H". The routine which converts g to "H" must also erase the voiced-consonant property of the g.

The letter x generates the sounds "K", "S" when followed by a vowel and "S" otherwise.

There are only three letters which do not require special handling: j (sound "H"), t ("T"), and y ("Y"). A summary of the major features of the Spanish algorithm is given in Fig. 5.

		Spanish	
a	AH2, AH2	n	N or M or NG
b	B or V	ñ	N, I3
c	K (S before e or i)	o	O1
ch	T, CH	p	P
d	D or THV	qu	K
e	EH1	r	R or R, R
f	F	s	S (Z before voiced cons.)
g	G (H before e or i)	t	T
h	silent	u	U
i	E1	v	B or V
j	H	x	S (K, S before vowel)
l	L	y	Y
ll	Y	z	S (Z before voiced cons.)
m	M (N at end of word)		

Sample: Buenos días. Yo puedo hablar español.

FIG. 5. The VS-6 sounds corresponding to the letters of the Spanish alphabet. The sample sentence says, "Good day. I can speak Spanish." See text for additional details.

When the end of one word and the beginning of the next word have the same vowel, as in "de escuela" or in "puerta abierta", one of the duplicate vowels is deleted. A convenient place to handle this is in the routine that is invoked when a space is encountered. If matching vowels are found on each side of the space, it is sufficient to back up the output buffer by one sound, thus deleting one of the vowels. This procedure does not disturb the stress pointers.

A useful reference is the linguistically-oriented 1960 edition of the Modern Language Association textbook, *Modern Spanish* (New York: Harcourt, Brace and World). Another useful book, both for sounds and for intonation, is *The Sounds of English and Spanish*, R. P. Stockwell, University of Chicago Press, Chicago (1965). A more formal and mathematical approach is given in *Spanish Phonology*, J. W. Harris, MIT Press, Cambridge, Mass. (1969). Early measurements of Spanish vowel durations were performed by Navarro: "Cantidad de las vocales acentuadas," *Revista de Filología Española* 3, 387 (1916), and "Cantidad de las vocales inacentuadas," *Revista de Filología Española* 4, 371 (1917).

Appendix B: Details of Italian

The Italian letters i, u, c, g, h, n, q, s, and z require special treatment. The vowels i and u are treated as they are in Spanish (except u before a stressed vowel is also taken as "w"). The letter c if followed by e or i is pronounced "ch" (Votrax "T", "CH"), and if ci is

followed by a vowel the *i* is silent. The sequence *cci* is treated as though it were *tci*. The letter *g* if followed by *e* or *i* is a "soft" *g* (Votrax "D", "J"), and if *gi* is followed by a vowel the *i* is silent. The sequence *gl* preceding an *i* is made with "L", "I3". (Some rather rare Italian words are produced incorrectly by this rule: in "glicine" and "negligente" the *g* is "hard".) The sequence *gn* is like the Spanish ñ. The *h* is silent. If *n* is followed by *b*, *m* or *p*, the *n* is pronounced "m", and this rule applies across word boundaries. For example, "un poco" is pronounced "umpoco". One easy way to handle this rule is to look at the preceding sound whenever *b*, *m* or *p* is encountered, and change a preceding "N" to "M". (There is some disagreement in the literature on this rule. In careful, formal speech, the *n* may be pronounced "N".) The *qu* is treated as "kw". The combination *sc* followed by *e* or *i* is "sh", and an *s* both preceded and followed by a vowel is pronounced "z". If *s* is followed by a voiced consonant (*b*, *d*, *g*, *l*, *m*, *n*, *r*, *v* or *z*), the *s* is pronounced "z". Either *z* or *zz* is pronounced "ts" (sometimes *z* and *zz* are pronounced "dz", but there is no simple rule for this). These rules are summarized in Fig. 6.

Italian			
a	AH2, AH2	m	M
b	B	n	N
c	K (T, CH before e or i)	o	O1 (ò is AW1)
cci	=tci	p	P
d	D	qu	K, W
e	A (è is EH)	r	R
f	F	s	S (or Z—see text)
g	G (D, J before e or i)	sc	before e or i SH
gli	L, I3, E1	t	T
gn	N, I3	u	U
h	silent	v	V
i	E1	z	T, S
l	L	zz	T, S

Sample: Buòn giorno. Posso parlare italiano.

FIG. 6. The VS-6 sounds corresponding to the letters of the Italian alphabet. The sample sentence says, "Good day. I can speak Italian." See text for additional details.

As in Spanish, when a word ends with the same vowel the next word starts with, one of the vowels is deleted (as in "amico onesto").

As in Spanish, the sound of *r* is produced incorrectly, due to the lack of a tapped or trilled *r* in the Votrax.

A useful reference is *The Sounds of English and Italian*, F. B. Agard and R. J. Di Pietro, University of Chicago Press (1965). Also useful is "Pronuncia e Grafia dell'Italiano", Amerindo Camilli, Sansoni, Florence (1947). These two books disagree on the pronunciation of unstressed *e* and *o*, and the simpler rule of Camilli has been implemented. Camilli states that unstressed *e* and *o* can be adequately rendered by the sounds of *é* and *ó*, and that the sounds of *è* and *ò* occur only in stressed syllables. Two papers on synthesis of Italian by M. Mezzalama & E. Rusconi may be found in "Speech Communication, Vol. 2, Speech Production and Synthesis by Rules", pp. 307-325, Ed. Gunnar Fant, Almqvist & Wiksell International, Stockholm (1975).

Appendix C: Details of Russian

Many Russian consonants when followed by certain vowels or the Russian "soft sign" become "palatalized". For example, the Latin transcription of the Russian word for "no" is "net", but this word is pronounced approximately as "nyet". This palatalization is handled by table entries showing whether a letter can be palatalized. When such a letter is encountered, another table entry is checked for the following letter to see whether palatalization is to be performed. If it is, a Votrax "I3" (essentially a very short "y") is added after the consonant. Figure 7 includes a list of palatalizable consonants and letters which trigger palatalization. (But the short "y" is not added in the case of a Russian "soft sign" occurring between consonants, since this would make an extra syllable.)

Vowels are assigned differing sounds depending on whether they are encountered before, at, or after the stress location. The three sounds are given in Fig. 7 for each vowel. Note that stressed vowels are given longer duration, which is a substitute for the loudness variation unavailable in the VS-6. Not shown in Fig. 7 is the fact that the letters "a" and "o" require special treatment in the syllable immediately preceding a stressed syllable, where both letters indicate VS-6 "AH1" (in other pre-stress and post-stress locations these letters are converted to VS-6 "UH1"). One way to handle this exception is to have a table of vowel sounds for the syllable immediately preceding stress (all entries in this table are the normal pre-stress values, except for "a" and "o"). If one keeps track of the last vowel encountered and its position in the output buffer, upon encountering a stressed vowel one can overwrite the output buffer position with the correct phoneme code, taken from the table of sounds for immediate pre-stress position.

Russian						
	а	UH1; AH	Р к	K; G	Р х	H, K, H
Р б	Б; P		Р л	L	ц	T, S; D, Z
Р в	V; F		Р м	M	Р ч	T, CH
Р г	G; K		Р н	N	ш	SH; ZH
Р д	D; T		о	UH1; O	Р ш	SH, T, CH
Y* е	E1; EH		Р п	P; B	ы	I1; I
Y* ё	AW1; AW		Р р	R	э	EH1; EH
ж	J, ZH; SH		Р с	S; Z	Y* ю	U1; U
Р з	Z; S		Р т	T; D	Y* я	E1; AH; I2
* и	E1; E		у	U1; U	ь	silent
й	Y1		Р ф	F; V	ь	I3

* Vowels which cause palatalization.

P Consonants which can be palatalized.

Y Vowels which require an initial Y1 if at the beginning of a word or after a vowel.

Prestress, stress and poststress (if different from prestress) variants of vowels are shown. Voiced/unvoiced variants of consonants are shown. Devoicing is caused by ц, ф, к, п, с, т, ш and punctuation. Voicing is caused by б, д, г, ж, з.

Sample: Дóбрый дéнь. Я гóворю по-ру́сски.

FIG. 7. VS-6 sounds for Russian. The sample sentence says, "Good day. I speak Russian".

A few Russian vowels are preceded by a short "y" sound (VS-6 "Y1") if the vowel comes at the beginning of a word or follows another vowel or the silent "hard sign". These special vowels are specified in Fig. 7.

Another context-dependent pronunciation rule is that in certain pairs of consonants the first consonant can be changed in its "voicing" characteristic. For example, the combination "vs" is pronounced "fs", and one says that the "unvoiced" s causes the preceding v to be "devoiced" to an f. Similarly, in the combination "kj" the voiced j (j as in the French word "jour") causes the preceding unvoiced k to be voiced to a g. Figure 7 shows which consonants cause voicing change and what are the changed sounds. These rules operate across word boundaries (at least if the words are pronounced all in one breath), so an intervening space is ignored in applying this consonant-pair rule. A punctuation mark (comma, period, etc.) also causes devoicing of a preceding (voiced) consonant. One way to implement these rules is to assign to each letter a "forcing" number (0 if non-forcing, 1 if it forces voicing, 2 if it forces devoicing), a "voicing" number (0 for letters which do not change, 1 for voiced consonants which can be devoiced, 2 for unvoiced consonants which can be voiced), and an alternate sound (the sound to be substituted if the voicing character is changed). Before assigning a sound to a consonant, its voicing property is compared with the forcing property of the following letter to see whether the alternate sound must be chosen.

When two unvoiced "stop" consonants (p, t, k) or two voiced stop consonants (b, d, g) follow each other, the VS-6 synthesizer deletes the first of the two consonants (unless a vowel precedes the pair). This causes no problems in English, but in Russian this makes such words as "kto" ("who") and "gde" ("where") lose the important initial sound. Here is a case of the synthesizer being more intelligent than is desired in the transitions it makes between sounds. The table of letter properties shows which consonants are stops, and when a pair of similarly-voiced stop consonants is preceded by something other than a vowel, a short pause (VS-6 "PA") is inserted between the two consonants.

Except for Fig. 7, the Russian alphabet has been avoided in this discussion in order to avoid complications for the typesetter. In the actual use of this algorithm on the PLATO system, a Russian set of characters is used.

Any standard Russian textbook can be consulted about the rules for the sounds: for example, see *Reading and Translating Contemporary Russian*, H. W. Dewey & J. Mersereau, Jr., Pitman Publishing Corp., New York (1963). A specific reference for intonation is *Zvuki i Intonatsii Russkoi Rechi*, E. A. Bryzgunova, Moscow (1969).

Articles by Bruce Arne Sherwood on speech synthesis

Sherwood, B. A. Fast text-to-speech algorithms for Esperanto, Spanish, Italian, Russian and English. *International Journal of Man-Machine Studies* 10, 669-692 (1978).

Sherwood, B. A. Schnelle textgesteuerte Sprachsynthese- Algorithmen für Esperanto, Spanisch, Italienisch, Russisch und Englisch. *Grundlagen-studien aus Kybernetik und Geistes-wissenschaft* 20, 109-121 (1979).

Sherwood, B. A. The use of synthetic speech in interactive computer systems. Invited seminar, Society for Information Display, Chicago (1979).

Sherwood, B. A. The computer speaks. *IEEE Spectrum* 16, 18-25 (1979).

Sherwood, B. A. Speech synthesis applied to language teaching. *Studies in Language Learning* 3, 171-181 (1981). Reprinted in *Esperanto in the Modern World*, ed. R. Eichholz and V. S. Eichholz, 431-449. Esperanto Press, Bailieboro, Ontario (1982).

Sherwood, B. A. New technology provides computer voices for education. *Speech Technology* 1, no. 1, 24-29 (1981).

Sherwood, B. A. Computer processing of Esperanto text. *Studies in Language Learning* 3, 145-155 (1981).

Cheng, C. C., & B. A. Sherwood. Technical aspects of computer-assisted instruction in Chinese. *Studies in Language Learning* 3, 156-170 (1981).

Sherwood, B. A. Komputila traktado de Esperanta teksto. In C. Bertin, ed.: *Unua Simpozio pri Komputiko*. Rennes, France (1981).

Sherwood, B. A. Spertoj pri sintezo de Esperanta parolado. In Helmar Frank, Yashovardhan, and Frank-Böhringer, ed.: *Lingvo-Kibernetiko*, 63-74. Gunter Narr Verlag, Tübingen (1982).

Sherwood, J., and Sherwood, B. A. Computer voices and ears furnish novel teaching options. *Speech Technology* 1, no. 3, 46-51 (1982).

Sherwood, B. A. Raporto pri Sintezo de Esperanta Parolado. Read at the INTERKOMPUTO-82 conference, Budapest (Dec. 1982).

New technology provides computer voices for education



Type 'n Talk by Votrax converts normal English spelling into synthesized speech. The device interfaces easily to most computers.



Speak and Spell by Texas Instruments uses compressed digital recording to provide spelling practice.

The natural man-machine interface of speech output is becoming part of today's computer based educational systems, as new digital technologies vie with older analog methods for acceptance.

Bruce Arne Sherwood

Assistant Director, Computer-based Education,
Research Laboratory and Department of
Physics, University of Illinois at
Urbana-Champaign

Speech, our most common means of communication, can augment computer-based text and graphics in the complex and difficult job of education. Possible uses of voice output include talking to young children who don't yet read, giving foreign-language dictation drills, and providing spoken commentary while a simulated comet orbits the sun for a physics student. Non-voice audio is used in teaching music and medicine.

Computer based education systems

require audio-output capabilities that challenge the state of present technology. Voice output used in computer-based education must be high quality, easy to create and deliver, inexpensive, and in some cases multilingual. Unfortunately, none of today's technologies fully meets all of these requirements. Three basic approaches [1] and their tradeoffs will be described:

- Analog recording.
- Compressed digital recording.
- Text-to-speech synthesis.



Bruce Arne Sherwood is Assistant Director and Co-chairperson of the Computer Based Education Research Laboratory and Associate Professor of Physics at the University of Illinois, at Urbana-Champaign. His research in the Plato project extends from system software, Plato-based physics teaching, multi-lingual text-to-speech synthesis, to problems in international communication. He has authored a textbook, "The TUTOR Language," and with his wife, Judith, has applied synthetic speech to teach languages. Sherwood received a BS in Engineering Science from Purdue, studied Physics at Padova, Italy under a Fulbright scholarship grant, and received a Ph.D. in Physics from the U. of Chicago.

Analog recording

Magnetic tape and disk recordings provide high fidelity not only of voice but of audio in general, including music. Computer-controlled tape cassettes with an addressing track permit selection of individual messages. However, the serial nature of the tape leads to access times of tens of seconds. This imposes severe constraints on lesson design, since only neighboring messages can be selected without long delays. Though unsuitable for computer-based systems that require branching flexibility, tape may be viable in some applications.

Figure 1 shows a random-access audio disk which overcomes the slow access time of tape. This device is manufactured by Education and Information Systems (see box for vendor addresses). This is a commercial version of a device originally designed for the University of Illinois PLATO system [2]. A 15-inch disk of audio-tape material holds 27 minutes of high-quality audio and provides access to any part of the disk in half a second.

Figure 2 shows the fast access mechanisms, consisting of high-speed rotational positioning of the disk and high-speed radial positioning of the playback head. Addressing requires 20 bits: 7 bits select one of 128 concentric tracks; 5 bits select one of 32 sectors, with the current sector identi-

fied by optical sensing of holes in the disk; and 8 bits specify how many sectors to play.

Once the playback head reaches the proper part of the disk, the head is lowered and a slow pinch wheel drives the disk under the head at audio-tape speed. During playback a 400-ms sector can be divided into finer slices by timing the system's audio-muting feature from the start of the sector. This muting feature makes it possible to start playing at the beginning of a measure of music, or at the beginning of a word inside a recorded sentence. Computer-controlled positioning is also used when recording from a microphone or other audio source.

Similarly, the audio portion of a videodisk could be used for analog audio output. The audio track of a videodisk offers 30 minutes of capacity per side. Worst-case access time is 5 seconds on commercial models but about 30 seconds on consumer models. Moving to a position within a few hundred frames of any current position can be accomplished in a fraction of a second because only a small mirror is moved.

Alternatively, entire video frames could contain analog audio, in which case buffering would be required to slow the signal down to normal speed. One such video frame could hold 5 to 10 seconds of audio, and a videodisk completely devoted to analog audio could provide a capacity of over 100 hours. Unlike audio tape and disk, videodisks cannot be recorded and reproduced locally, since this requires special equipment.

Audio disks have been used in the PLATO system in reading, veterinary medicine, and languages. Reading lessons for young children used audio disks to talk to the child about letters displayed on the terminal screen [3]. In veterinary medicine, sounds of diseased hearts were collected in the clinic and gathered onto an audio disk to provide practice in identifying heart problems [4]. This non-voice use could not be handled by the voice-oriented digital devices to be discussed later.

Many foreign-language lessons rely on audio disks to provide high-fidelity native speech under computer control [5]. The audio disk offers a unique capability for the student to record his or her own voice, permitting comparison with the native speaker by quickly jumping between the two messages. A single track is sufficient for the student's recordings. Providing the same facility with tape would be awkward: space must be left after every message, and switching between messages is slow.

Student recording is impossible with videodisk.

Compressed digital recording

Compressed digital recording (CDR) is a technique for greatly reducing the storage requirements for digitized speech. If the voice is merely sampled by an analog-to-digital converter, the resulting samples may require as much as 50,000 bits to represent one second of speech with good fidelity. This volume of data is inconveniently large for storage, transmission, and management within a computer-based system. Various techniques, the most popular being linear predictive coding (LPC), can greatly reduce the data volume by taking into account the special nature of the speech signal [6]. (See page 17 article by Wakita.) Such techniques depend on the fact that the mouth organs change position rather slowly. Underlying the rapidly changing time-domain waveform are slowly varying frequency-domain parameters that correspond to human vocal-tract resonances. At present it is possible to compress speech to about 1200 bits/s and still retain quality sufficient for many purposes.

Unlike analog recording, CDR as presently implemented is useful only for voice, not for other audio output such as music or heart sounds. As in analog audio, a few addressing and duration bits are sufficient to call up a compressed digital message. Playback involves passing the digital data through a voice player which is a digital signal processor.

These CDR players have often been called "synthesizers," but this has led to a great deal of confusion, since they merely play back a (compressed digital) recording. It is the author's choice to reserve the word "synthesizer" solely for text-to-speech systems. Otherwise we should call record or tape players "synthesizers," since they transform vinyl or magnetic-domain variations into speech or music, just as CDR players transform into speech electronic states recorded in silicon.

The major use of CDR is in devices with a small fixed vocabulary of words and phrases, such as talking calculators for the blind. CDR has also been used to represent individual speech sounds or even whole words for text-to-speech synthesis.

Perhaps the best-known educational CDR product is Texas Instruments' "Speak and Spell," (see page 24). It sells for about \$50 and gives practice in spelling 200 English words. Each word in the vocabulary as compressed to 1200 bits/s for storage in read-only memory. Creating this vocabu-

lary was very costly and involved considerable hand editing of the compressed digital data in order to optimize the speech quality.

At somewhat higher bit rates CDR can be an automatic process, just like analog recording. However, at the lowest usable bit rates the recording process is labor-intensive. Compressing one word digitally can cost as much as \$100! Such costs can be afforded for a mass-produced device, but this is prohibitively expensive for generalized computer-based education. Some manufacturers of CDR players offer standard vocabularies at low cost for popular uses (e.g., numbers, names of letters, etc.). When standard vocabularies suffice, CDR can be inexpensive.

Centigram Corporation has a new voice development system, a CDR "workbench" with realtime automatic compression to 4800 bits/s and interactive software tools for manipulating the recorded speech (e.g., clipping off the beginning or end of a message, filing the data by message name, etc.). While the approximate \$30,000 price of this system will make in-house compressed digital recording feasible for some applications, the Centigram "Lisa" player is presently expensive (about \$2500 in single quantities, though there are plans to produce an inexpensive player). What is needed is a combination of low cost in-house recording and low cost CDR players.

The MISS system (Micro-programmed Intoned Speech Synthesizer) at Stanford University performs nearly automatic compression to 4800 bits/s, providing good quality speech for computer-based educational applications [7, 8]. For highest quality, about one percent of the words require pitch adjustments by hand. At Stanford, MISS has been used to add voice explanations, commentary, and emphasis in mathematics courses.

MISS has also been used in an Armenian language course. Armenian sentences are prerecorded, but the English explanations are produced by concatenating LPC-coded words. Text-to-speech rules manipulate critical pitch and duration aspects of these words to produce appropriate rhythm and intonation. Given a large vocabulary of words, the generative capabilities of such a system could approach that of purely synthetic text-to-speech while preserving the high quality and intelligibility of human speech. More work is needed to improve the smoothness of transitions from one word to the next, but the approach is intriguing. This scheme is not applicable to analog recordings, since the concatenation of

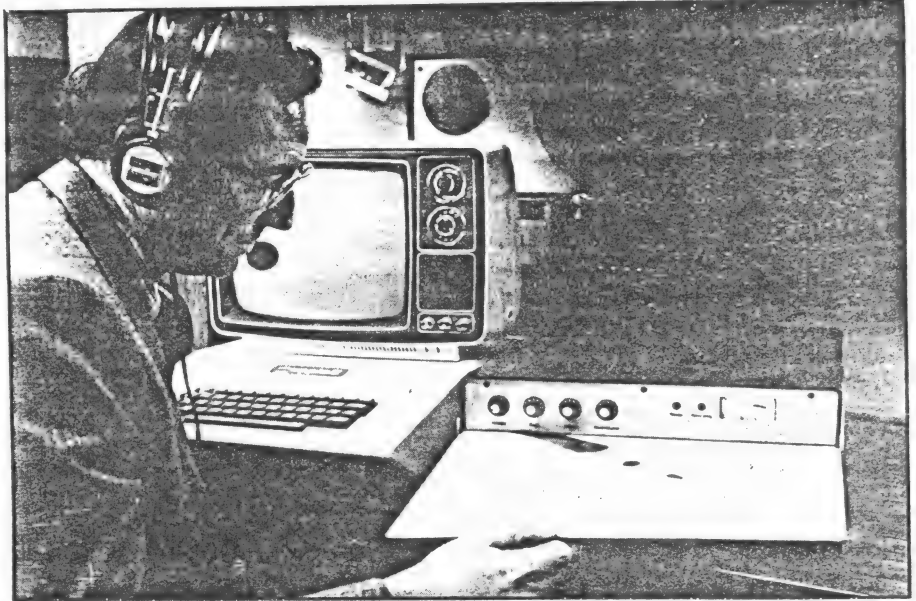


Figure 1. EIS random-access analog audio disk is made of audio-tape material and enclosed in a protective envelope.

unaltered individual words produces unnatural speech, and the access time for existing analog players would insert abnormally long pauses between words.

This semi-synthetic speech has been used at Stanford as an authoring tool during curriculum development. In the initial phases, the author needs easily modified audio but it need not be of the highest quality. Only after the educational materials have been through the first cycles of revision are the final forms of the messages recorded as complete phrases, to get the best speech quality. If an author includes the phrase "Answer this question" in a lesson, the system can produce the corresponding speech in two ways. If a full measure with the title "Answer this question" has been recorded, this will be played. If only the individual words have been recorded, these words are concatenated.

A new version of this system is being used by the Computer Curriculum Corporation in an elementary reading course for early primary grades and will be used in a 6-year English curriculum for speakers of Spanish or Chinese. Both straight recordings and word-concatenation techniques are used in these materials. Thirty to fifty seconds of computer analysis are required to compress one second of speech, and, as before, about one percent of the words require some manual adjustment of pitch. Male speech is represented by about 3300 bits/s, female speech by about 4400 bits/s, and children's speech or singing (both of which require some extra hand editing) by about 5000 bits/s. Clusters of 16 terminals

are provided audio by a nearby CDR player with a 35-megabyte hard disk of voice data (equivalent to 2.5 hours of 4000 bit/s speech). The player can multiplex six simultaneous voices among the 16 users.

Although the quality of word-concatenated speech is not as high as is desirable, the multiple uses of a recorded word in different contexts saves a great deal of space. The sheer volume of voice data involved in these audio-oriented courses requires making a trade-off between quality on the one hand and storage space, recording time, and LPC-analysis time on the other. Fully recorded messages are used when the highest quality is essential, while word-concatenated speech is used wherever educationally possible. Almost two hours of speech data for elementary reading have been created, and the English language course will eventually require eight hours of speech data.

James Parry at the University of Arizona has been working to automate compression at the 1000 bit/s level, with a view toward being able to store CDR speech centrally in the PLATO system. Terminals run at 1200 bits/s, which would permit sending CDR data to an inexpensive player attached to the terminal. Such transmission can temporarily saturate the terminal line, unlike the low bit rates required to select analog or CDR messages stored at the terminal, but central storage simplifies distribution and updating problems.

Because there are many different kinds of voice-output devices in use, Parry looks to software tools to provide some device

independence. Generalizing on the MISS scheme which can output a phrase in two different ways, Parry has created a set of routines to manage a database of titles and messages, with matching output routines to drive various devices appropriately. Assuming the messages are available on an analog recording, in a CDR data bank, or can be synthesized from the title itself, one statement in a computer-based lesson could drive whatever output device is being used at the terminal.

In addition to the complexity and expense of the recording process, a problem for some applications is that at the lowest bit rates, CDR-speech quality may not be adequate for some purposes. The quality is likely to improve greatly, given the large industrial investments in CDR intended for various consumer products.

Eventually, standards must emerge for CDR at various bit rates. Then, recordings made with one analysis system could be played on many different CDR players. Without such standards CDR cannot become generally useful.

Text-to-speech synthesis

The conversion of text into speech begins with an algorithm to determine the sounds, rhythm (duration of sounds), loudness contour, and pitch contour (intonation, and tones for languages such as Chinese). The corresponding speech signal is synthesized from these parameters. The basic advantage of text-to-speech synthesis is that new messages can be generated with unlimited vocabulary, which removes the limitations of prerecorded analog or digital messages. This has important benefits for educational applications. The major disadvantage is that the quality of commercially-available systems is inadequate for many educational purposes. However, some laboratory-synthesis systems already produce good speech.

A good text-to-speech algorithm must include not only spelling rules (and exceptions) but also phonological rules which specify how sounds are modified in context. For example, "did" and "you" when by themselves are pronounced "did" and "yoo," but together they are usually pronounced "dijoo."

Syntax and semantics can also be important. In English, syntactical analysis is necessary to distinguish between such words as the noun "PROject" and the verb "proJ-ECT." Semantic analysis is needed to distinguish between the nouns "bow" ("bou," the front of a ship) and "bow" ("boe," a knotted ribbon) or between the present and past tenses of the verb "read" (pro-

nounced "reed" and "red").

Some languages are simpler. An extremely simple algorithm suffices for the constructed language Esperanto. Each letter has only one sound that is little modified by context; stress lies always on the next-to-the-last vowel in a word; and the rhythm is simple (all unstressed vowels have the same duration and all stressed vowels are about 50% longer). Although Spanish requires many more rules than Esperanto, there are essentially no exceptions, so once the algorithm has been worked out it will give reasonable results for all text.

Most languages seem to fall between Esperanto and English in complexity for text-to-speech synthesis [9, 10, 11, 12, 13]. Since there are many possible human readings of a text, there will always be room for improvement or change in text-to-speech rules. However, for most purposes it is adequate to approximate one possible human reading.

In addition to handling spelling problems, algorithms for converting English text to high-quality speech must deal with subtleties of the English sound system, including its intricate rhythm and stress patterns. However, there are algorithms [14] which can run in real-time on microcomputers and pronounce English adequately to be

helpful to the blind. Such algorithms have five hundred to a thousand rules and exceptions (e.g., the "silent e" rule, with "have" listed as an exception). This is the basis for talking computer terminals [15] (Maryland Computer Services and Triformation Systems), for talking computers [16], and for a reading machine using optical character recognition (Kurzweil Computer Products). These algorithms make mistakes such as pronouncing "live" as "lie v" even when it should be pronounced "liv," but the usefulness to the blind person overwhelms these relatively minor errors.

Many educational applications demand essentially error-free English synthesis. Such quality requires a large pronouncing dictionary and a complex program [17] (see the article "Linguistic-based algorithms offer practical text-to-speech systems" on page 12) that may not run in real time unless some form of phonetic text is used. In systems with insufficient processing resources to handle standard text, a compromise solution would be to use an imperfect but inexpensive algorithm and have the author of a computer-based lesson use abnormal spellings such as "liv" where necessary when writing the lesson.

The PLATO system uses a simple phonetic spelling scheme called "World English Spelling" for adding English-voice

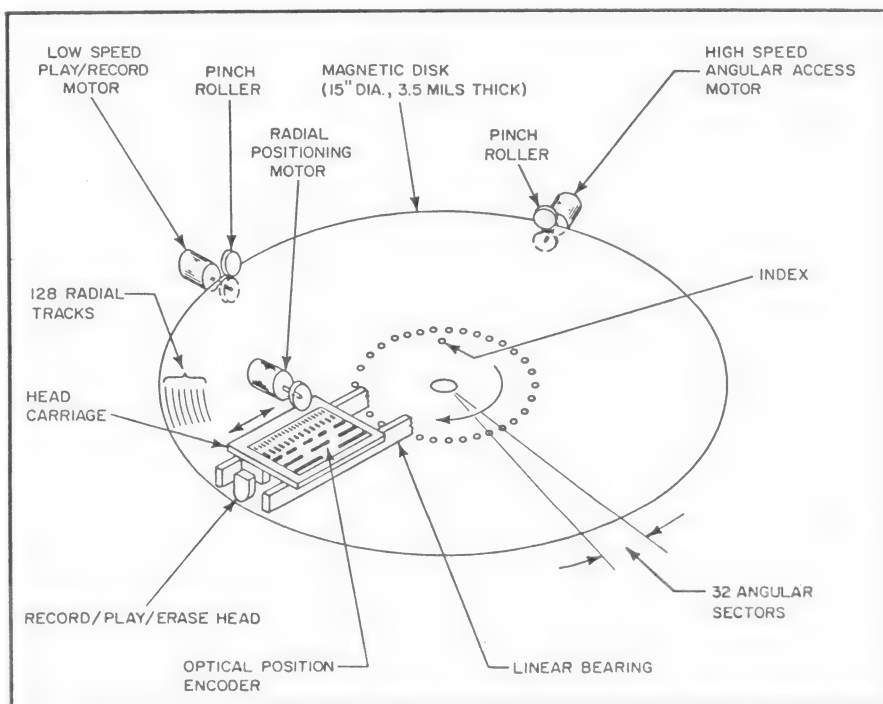


Figure 2. Two motors give high-speed access in an EIS disk. Fast angular-access and radial positioning motors reach any part of the 27-minute disk in a half-second or less. Once in position, the slow play/record motor rotates the disk at 5 rpm—corresponding to about cm/s (3 in/s) under the play/record/erase head.

output to lessons. An interactive editor program immediately speaks the entered text to the author for checking. Options are also available for Esperanto, Spanish, and the International Phonetic Alphabet.

Several types of mechanisms can produce voice output from text-to-speech algorithms. These may be characterized as "phonemic synthesizers," "allophonic synthesizers," and "full synthesizers." A phonemic synthesizer for English has a stock of about forty basic sounds or "phonemes" (12 vowels including ah, eh, ee, er, oo, uh, etc.; 3 diphthongs ie, oi, ou; and 24 consonants including p, t, k, r, etc.). Many of these phonemes are actually families of similar sounds called "allophones," and a device with perhaps 120 of these is an allophonic synthesizer. For example, the phoneme /l/ has two quite different allophones at the beginning and end of English words. Perhaps you can feel the difference in tongue placement for the /l/ in "let" and in "tell." There is no clear-cut distinction between phonemic and allophonic synthesizers, because the allophonic variation could be built into the device. In that case, the sound produced in response to the algorithm's request for a phoneme could depend on context.

The complete text-to-speech algorithm can be built into the device, in which case one has not a phonemic or allophonic synthesizer but a full text-to-speech synthesizer, to which one can simply send English-character strings to produce speech. However, the increasing convenience in going from allophonic, to a phonemic, to a full synthesizer is paid for in decreasing control over the speech output.

In particular, it would be inconvenient or even impossible to make a full English synthesizer produce good Spanish (by spelling Spanish according to English spelling rules). Too many of the language-dependent decisions are made inside the device. For example, Spanish does not share with English the allophonic variation of /l/ at the beginning and end of words.

Typically, a full synthesizer is constructed out of a hardware phonemic or allophonic synthesizer plus text-to-speech software. Hence, it may be possible to bypass this software when other languages are to be synthesized, using an external text-to-speech algorithm. The allophones of one language are not quite right for another language, but one can approximately synthesize speech for a different language if the available stock of allophones is large.

Examples of commercial synthesizers are the Votrax line of products including phonemic, allophonic, and full synthesizers.

An English text-to-speech algorithm is built into the Votrax "Type 'n Talk" device (see page 24). It accepts standard English-character strings but also permits input of phonetic text for foreign words.

Allophonic synthesizers are made by Texas Instruments [18] and by Street Electronics (the Echo II synthesizer). Both of these are based on Texas Instrument's LPC player chip, with allophones extracted from LPC-coded human speech (similar to the MISS word-concatenation scheme). Both are available with programs to convert English text into allophones. The allophones are stored as data, so different allophone sets could conceivably provide a variety of voices (e.g., male and female) or non-English speech. The English algorithms available for all of these devices make mistakes of the liv/live variety, and the programmer must occasionally "misspell" in order to get the desired results.

Words and allophones are not the only speech elements which can be concatenated for text-to-speech synthesis. For example, work in our laboratory [19] is taking another look at "diphone" synthesis. Two-phoneme sequences are extracted from human speech by pitch-synchronous short-time Fourier-transform techniques. These diphones capture much allophonic variation automatically, and they contain the complex, rapid transitions between phonemes which carry much of the speech information.

The unusual characteristics of text-to-speech synthesis, especially the capability to generate new messages by program, have been investigated in our laboratory in the area of computer-based language teaching.

Comparisons

For non-speech applications such as heart sounds only analog recording is usable at present. (Except for music synthesis.) For voice output the high quality of analog recording is matched only by high-bit-rate CDR (above approximately 5000 bits/s). Currently available text-to-speech and low-rate CDR systems do not supply adequate speech quality for many educational applications. This may change soon because so much industrial and academic research is aimed at improving digital speech. Some laboratory text-to-speech systems are already capable of high quality. Let's assume that adequate quality will eventually be available, and compare the advantages and disadvantages of the several methods.

A major disadvantage of analog recording is that it is a "live" medium that requires

studio facilities and a good reader. Creation and editing of analog recordings are awkward. For example, at the University of Illinois Language Learning Laboratory, audio disks are prepared by a multistep process. A script is read in a special studio and recorded on high-fidelity tape. Then, approximate message marks are added by hand to another track of the tape. Next, this marked tape is used to record several audio disks in parallel. Many database manipulations are performed to associate track, sector, and duration specifications with each message to make sure the messages start and stop precisely where desired.

The newer version of the audio-disk device, with its higher fidelity and more precise accessing, might encourage direct recording on the disk, to make the original master. However, there would still be problems with packing the disk efficiently and managing the database of messages, in addition to the studio aspects.

After a disk is made, modification can be difficult. Messages cannot be lengthened in place unless extra space is left at the end of each original message. In a time-sharing environment, analog audio cannot be stored centrally and transmitted over the terminal-communication lines. This causes distribution and update problems when revisions are desired.

CDR shares many of the problems of analog recording. However, at low bit rates it has the advantage that it can be stored centrally and transmitted from a time-sharing system to terminals for reproduction on local players. In a microcomputer environment, CDR data could reside on floppy discs along with programs, thus sharing some of the standard hardware. At 4000 bits/s, 30 minutes of CDR speech would occupy a megabyte of storage, requiring a full double-sided double-density floppy disk. This amount of data is large compared with computer-based education programs that typically require about 32-kbytes for 30 to 60 minutes of instruction.

A large fraction of program and data storage could be consumed by CDR data, whether on floppy or time-sharing disks. In one CDR application, the economics of storage and analysis forced trade-offs between recordings and word-concatenation.

CDR speech offers editing advantages over analog recordings because the data can be moved around in computer memory. However, the editing process in both cases involves rather clumsy and time-consuming manipulations that are unnatural in the computer environment. One must listen to a message, then move forward or

back to include or exclude data as needed. As pointed out by Levine and Sanders [7], revising analog or CDR recordings can be impossible without the original reader, unless a change of voice can be tolerated in some messages. This is particularly serious for word-concatenation schemes.

In contrast, creation, editing and distribution of the text used in text-to-speech synthesis is not a "live" process and does not require a special studio or reader. There are highly developed software tools for manipulating text in computer systems. Storage space and transfer rates are small (about 100 bits/s of speech). Even if all the text in an educational program were spoken, the program would be just the normal size.

Of critical significance is the capability not merely to retrieve pre-recorded mes-

sages but to generate new messages by the algorithmic creation of character strings. This makes text-to-speech a much more general output scheme than either analog or digital recording. In the long run, this aspect is likely to make text-to-speech the preferred approach in computer-based education if the quality ever reaches adequately high levels at adequately low costs.

Many instructional applications require voice output. The use of audio in mathematics at Stanford is perhaps the only major application of voice output where audio wasn't absolutely required but was used to heighten the interaction with the student. In the long run, such optional audio will be used by authors of computer-based materials only if it is very easy to do. This too tends to favor text-to-speech output

because of the ease of editing and revision.

In the last couple of years the cost of text-to-speech synthesis has dropped to one tenth thanks to new chips and to competition, but the speech quality has stayed nearly constant (and too low for many purposes). It is hoped that the next few years will see increased effort to raise the quality while holding the costs constant. Until the quality reaches appropriate levels, analog recording and high rate CDR must continue to be used in the many applications that require high quality speech.



References

1. **Sherwood, B.A.** (1979). The computer speaks. *IEEE Spectrum* 16, 18-25.
2. **Smith, S., & Sherwood, B.A.** (1976). Educational uses of the PLATO computer system. *Science* 192, 344-352. Reprinted in *Electronics: the Continuing Revolution*, edited by Abelson, P.H., & Hammond, A.L. American Association for the Advancement of Science. Washington, D.C. (1977). PLATO is a registered trademark of Control Data Corporation.
3. **Yeager, R.** (1976). Using audio with CAI: experiences of the PLATO elementary reading project. *Proceedings of the Association for the Development of Computer-based Instructional Systems (ADCIS)*, August 1976, p. 227 (Minneapolis). ED 128 005.
4. **Muselman, E.E., & Grimes, G.M.** (1976). Teaching recognition of normal and abnormal heart sounds using computer-assisted instruction. *Journal of Veterinary Medical Education*, 3, 9-12.
5. **Marty, F., & Myers, M.K.** (1975). Computerized instruction in second-language acquisition. *Studies in Language Learning*, 1, 1.
6. **Rabiner, L.R., & Schafer, R.W.** (1978). *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, N.J.
7. **Levine, A., & Saunders, W.** (1978). The MISS speech synthesis system. Technical Report No. 299, Institute for Mathematical Studies in the Social Sciences, Stanford University.
8. **Sanders, W.R., Levine, A., & Laddaga, R.** (1979). The use of MISS in computer-based instruction. *Proceedings of the Association for the Development of Computer-based Instructional Systems (ADCIS)*, February 1979.
9. **Lesmo, L., Mezzalana, M., & Torasso, P.** (1978). A text-to-speech translation system for Italian. *International Journal of Man-Machine Systems* 10, 569-591.
10. **Sherwood, B.A.** (1978). Fast text-to-speech algorithms for Esperanto, Spanish, Italian, Russian, and English. *International Journal of Man-Machine Studies* 10, 669-692.
11. **Carlson, R., Galyas, K., Granström, B., Petersson, M., & Zachrisson, G.** (1980). Speech synthesis for the non-vocal in training and communication. *Speech Technology Laboratory Quarterly Progress and Status Report* 1/1980, pp. 14-28. Royal Institute of Technology, Stockholm.
12. **Maggs, P., & Trescases, P.** (1980). De l'écrit à l'oral: un programme sur micro-ordinateur pour machine à parler à l'usage des aveugles francophones [From the written to the oral: a micro-computer program for a talking machine for use by blind speakers of French]. *La Française Moderne*, 48 no.3, 225-245.
13. **Cheng, C.C., & Sherwood, B.A.** (1981). Technical aspects of computer-assisted instruction in Chinese. *Studies in Language Learning* 3, 156-170.
14. **Elovitz, H.S., Johnson, R., McHugh, A., & Shore, J.E.** (1976). Letter-to-sound rules for automatic translation of English text to phonetics. *IEEE Trans. on Acoustics, Speech, and Signal Processing* 24, 446-459.
15. **Kutsch, J.A.** (1976). A talking computer terminal. Report 76-815 of the Dept. of Computer Science, University of Illinois at Urbana-Champaign.
16. **Maggs, P.** (1981). A multilingual talking computer for the blind and speech-handicapped users. 14th Hawaii International Conference on Systems Sciences 2, 312-324.
17. **Allen, J.** (1976). Synthesis of speech from unrestricted text. *Proc. IEEE* 64, 433-442.
18. **Lin, K-S., Goudie, K.M., Frantz, G.A., & Brantingham, G.L.** (1981b). Text-to-speech using LPC allophone stringing. *IEEE Trans. Consumer Electronics* CE-27, 144-152.
19. **Glinski, S.** (1981) Diphone speech synthesis based on pitch-adaptive short-time Fourier transform. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Vendors of commercial products described in this article

Centigram Corp.
155A Moffet Park Dr.
Sunnyvale, CA 94086 (408-734-3222)

Computer Curriculum Corporation
1070 Arastradero Road
Box 10080
Palo Alto, CA 94303 (415-494-8450)

Control Data Corporation
Box O
Minneapolis, MN 55440 (612-853-4541)

Education and Information Systems, Inc.
804 N. Neil
Champaign, IL 61820 (217-352-4252)

Kurzweil Computer Products, Inc.
33 Cambridge Parkway
Cambridge, MA 02142 (617-864-4700)

Maryland Computer Services
2010 Rock Spring Road
Forest Hill, MD 21050 (301-879-3366)

Street Electronics Corporation
3152 E. La Palma Ave., Suite C
Anaheim, CA 92806 (714-632-9950)

Texas Instruments
Speech Technology Center
P.O. Box 6448, M.S. 3036
Midland, TX 79701 (915-685-6632)

Triformation Systems
3132 S.E. Jay St.
Stuart, FL 33494 (305-283-4817)

Votrax
500 Stephenson Highway
Troy, MI 48081 (313-588-0341)

Computer Voices and Ears Furnish Novel Teaching Options

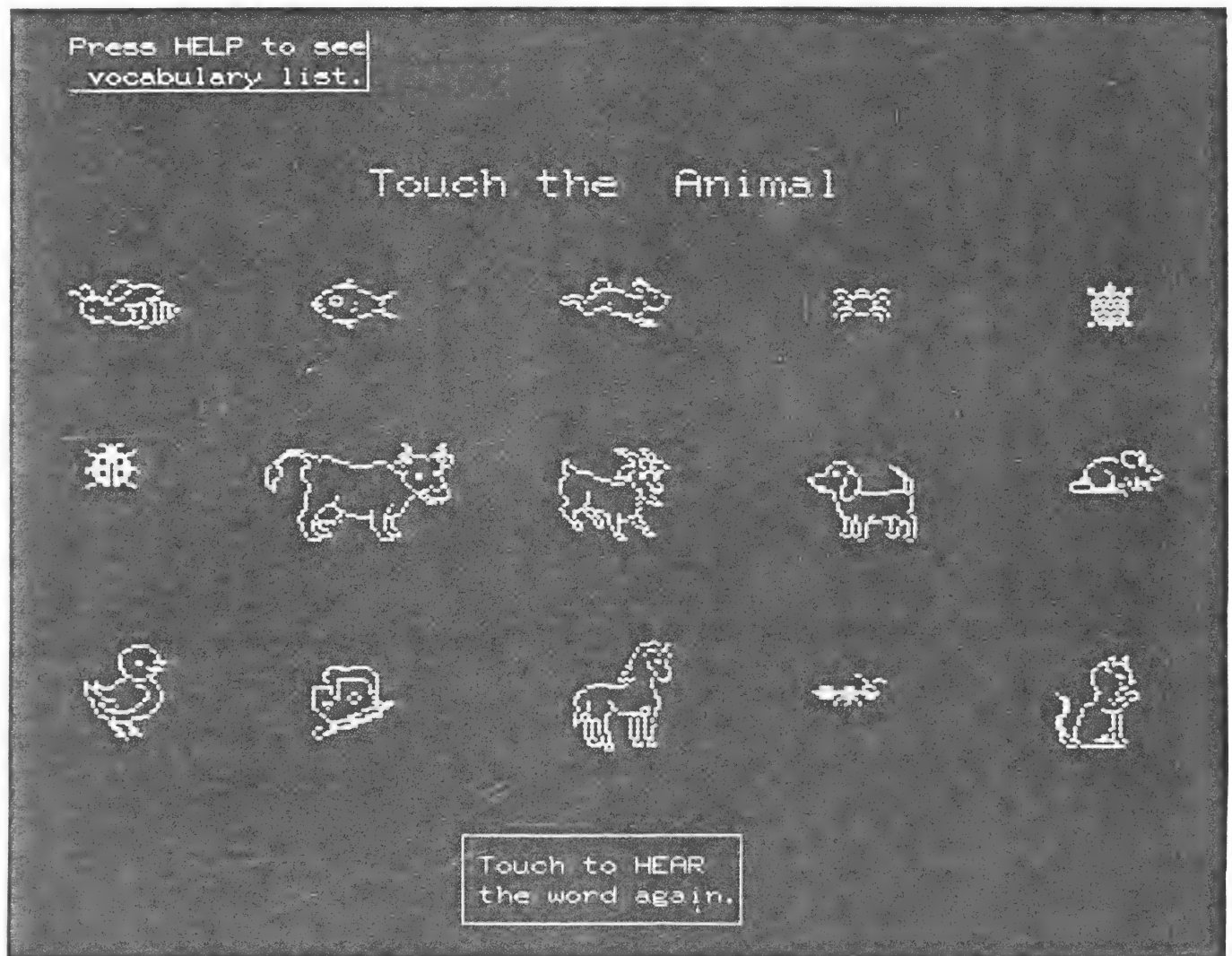


Fig. 1

*At least one language, Esperanto, can now be taught with
the aid of speech recognition and synthesis.*

*The student hears the name of an animal and has to touch the correct figure.
Touching the box at the bottom of the screen lets the student hear the word again.*

Judith N. Sherwood

Specialist in Automated Education

Bruce Arne Sherwood

Assistant Director

Computer-Based Education
Research Laboratory

University of Illinois at Urbana-Champaign

SPEECH TECHNOLOGY is opening up new possibilities for computer-based language instruction using both word recognition and text-to-speech synthesis. But caution is required. Students ought to hear nativelike speech, but commercially available speech synthesizers have not provided adequate quality. However, laboratory systems are already capable of producing good quality speech. When such systems become commercially available, they will be useful in teaching foreign languages.

In anticipation of such devices, we have experimented with the use of text-to-speech synthesis in the computer-based teaching of Esperanto. Esperanto has a rather simple system of sounds and can be synthesized adequately for many instructional purposes. Of course, computer-based instruction using synthetic speech in the student's own native language has been possible for some time [1], since the quality can be substantially less.

Part of a linguistics course at the University of Illinois involves the study of Esperanto [2] (see "Esperanto: An Overview"). One of the instructional tools is a set of interactive Plato lessons [3]. (The Plato system is a development of the University of Illinois and also a product of Control Data Corp.) Some of these lessons exploit the capabilities of synthesized Esperanto [4, 5, 6] in order to speak to the student. For example, Fig. 1 shows a screen presentation in which the student sees pictures of animals. The student hears a synthesized Esperanto name of one of the animals and must touch the correct picture.

Figure 2 shows one of the beginning lessons in the series, on sounds and spelling (which we and Michael Pogue wrote). The synthesizer says the word "mini" and the student must touch the correct box. If the student incorrectly touches the box labeled "mene," a synthesized audio feedback is immediately produced from the text "nat mene, mini" (pronounced "naht MEH-neh, MEE-nee").

On-Line Speech Synthesis

A critical advantage of speech synthesis in computer-based education is the ability to generate a message as needed. Figure 3 shows, for example, another part of the beginning lesson, involving dictation of single words. The synthesizer said "semi," but the student typed "seme" (inappropriately using English spelling). Using the student's own response, the program immediately generates the text "nat seme, semi" (pronounced "naht seh-

The student hears the word "mini" and must touch the box that contains that word. If the box containing "mene" is touched (incorrectly), a response is constructed from the text "nat mene, mini" (heard as "naht MEH-neh, MEE-nee").

The word "semi" was said, but the student incorrectly typed "seme." The program, using real-time text-to-speech synthesis, generates the response "nat seme, semi" (pronounced "naht seh-MEH-eh, SEH-mee").

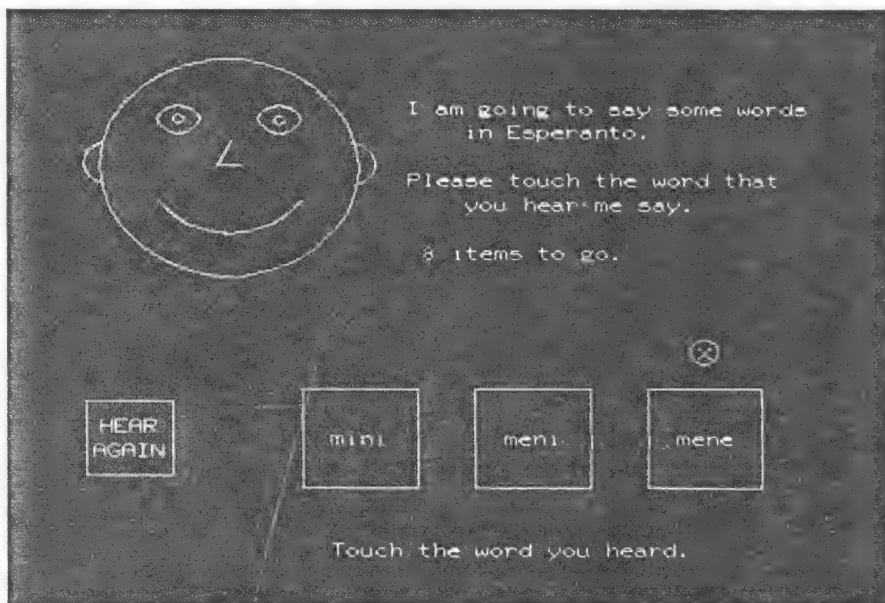


Fig. 2

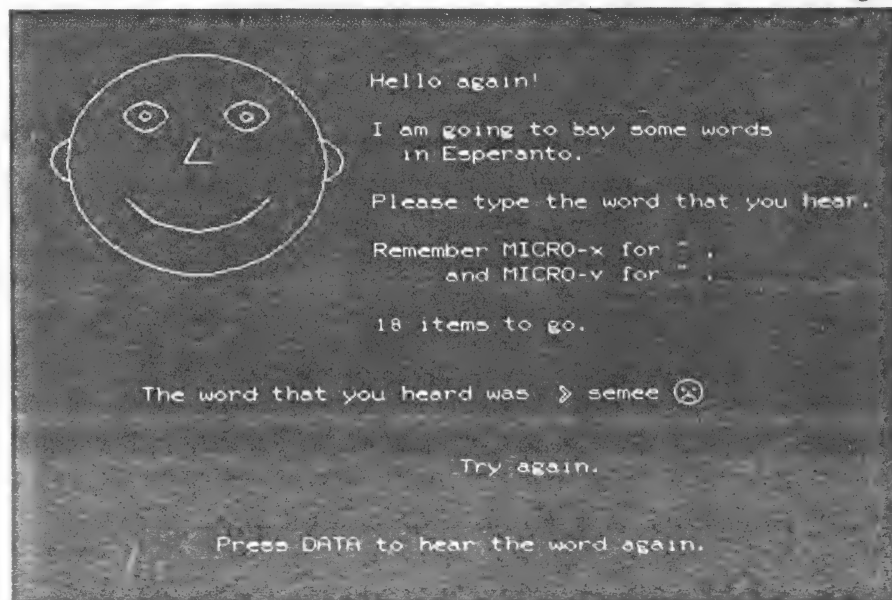


Fig. 3

MEH-eh, SEH-mee"). In this way the student immediately hears the contrast between his or her word and the correct one. This kind of presentation is possible only with text-to-speech synthesis, because it is impossible to anticipate and record all the possible wrong responses a student might make.

Another example of this unique algorithmic capability is seen in a data-base manipulator [3] that allows an instructor to

prepare a vocabulary exercise simply by typing words and definitions without doing any programming. After the instructor has entered the data, the student can choose among various drill options: see a word or definition and touch the corresponding definition or word, hear the word and type it, or hear the word and touch it or its definition. The audio options depend on the ability to convert text stored in the data base into speech (currently

implemented for Esperanto and Spanish). The same textual data can be displayed on the screen or synthesized into speech.

Figure 4 shows a more ambitious dictation exercise [3]: the student hears an entire sentence and must construct it by touching the appropriate words. The form of the presentation teaches something about word building, because the student adds endings to form certain words.

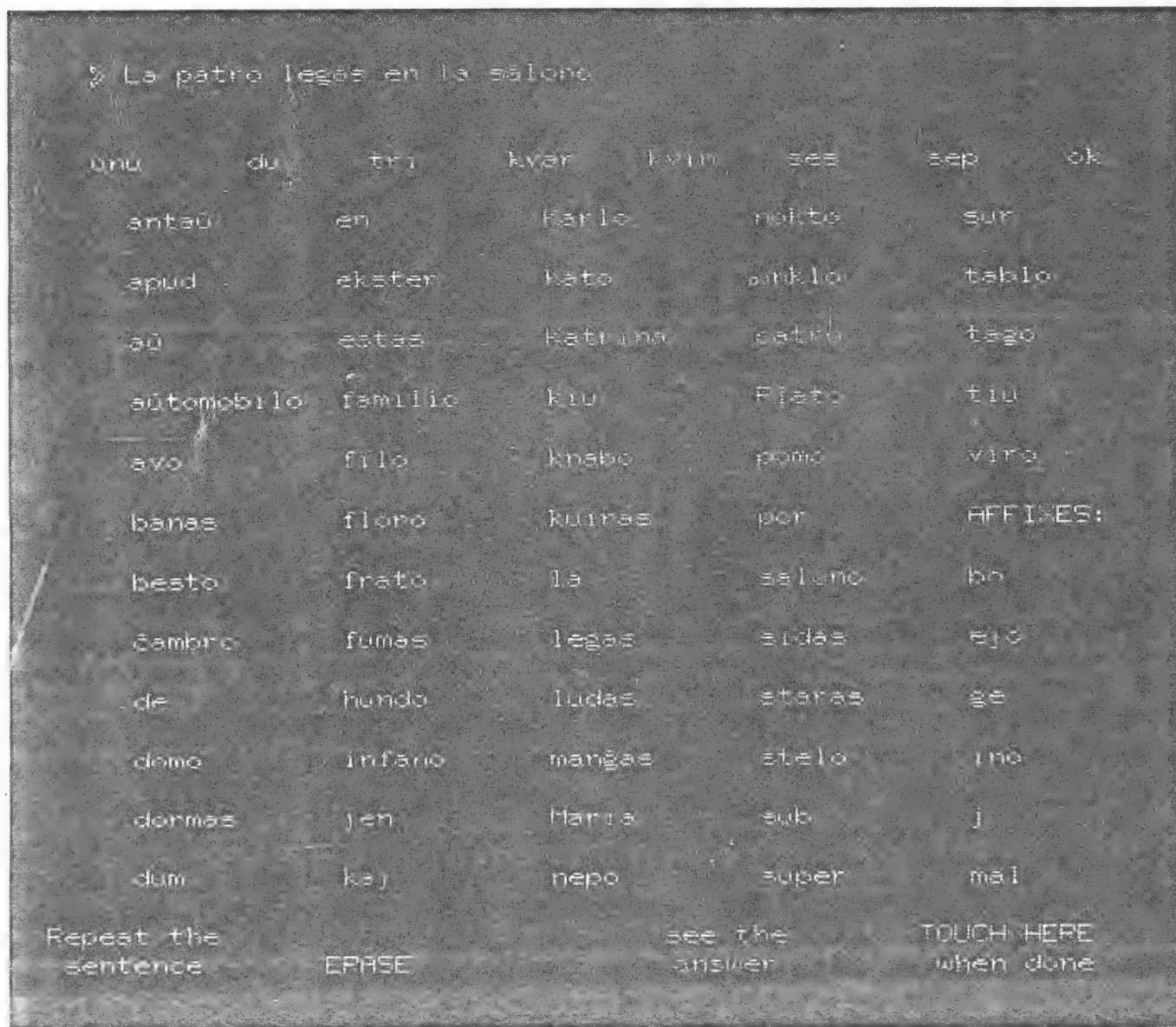


Fig. 4

The student hears an entire Esperanto sentence and has to construct it by touching the words and affixes on the screen. At the top of the screen one sees the student's partial answer.

The text-to-speech conversions performed in these applications are done in a few milliseconds in the central CDC Cyber computer: a second of speech requires about 1.4 ms of processing time. The codes corresponding to the text (sounds, durations, and pitches) are transmitted at 1200 bits/s to the terminal, which has a Votrax VS-6 synthesizer attached. A typical sentence requires about 50 bytes, which takes less than half a second to reach the terminal and start being spoken. The net effect is that a response to a student's input is heard within a half second. At present the entire sentence is sent to a buffer in the synthesizer before it can begin to be vocalized, but this scheme could be changed to permit the voice to start as soon as the first phoneme codes arrive, yielding a response time of 0.1 to 0.2 s.

Better Synthesis Needed

Speech synthesis opens up new kinds of pedagogical techniques in computer-based teaching, but to be fully satisfactory, the quality of the synthetic speech must be improved. That is likely to occur in the next few years. For Esperanto the quality is already adequate where there is sufficient context. For example, students understood the synthesized speech sufficiently well in situations with lots of context, as in multiple-choice drills like those of Figs. 1 and 2 and in the data-base-driven vocabulary drills. On the other hand, students had noticeable problems in situations such as in Fig. 3 (dictation of isolated words without any context or hints on the screen) and Fig. 4 (dictation of a long sentence). The basic problem is that the Votrax VS-6 synthesizer does not produce some consonants very well, so that certain distinctions among consonants can be difficult. For an experienced Esperanto speaker the synthetic voice is readily understandable, but for a beginner problems occur in some situations.

Except for the Stanford University and Computer Curriculum Corporation semi-synthesis work described earlier [1], perhaps the only other application of text-to-speech synthesis in language instruction is in phonetic Spanish reading lessons for native-Spanish-speaking children at the Gabe Allen Elementary School in Dallas, Texas. In order to obtain the highest possible quality from the Votrax ML-1 allophonic synthesizers used in this work, the Dallas group employed finely tuned phonetic text rather than standard



JUDITH N. SHERWOOD is a specialist in automated education at the Computer-Based Education Research Laboratory of the University of Illinois at Urbana-Champaign. She writes on-line documentation of the Plato system, consults with authors of computer-based materials, and is responsible for several system utilities. She has written a number of computer-based lessons to teach Esperanto and has experimented with using text-to-speech synthesis in language teaching. Sherwood received a BS in statistics from Purdue and an MS in statistics from the University of Chicago.



BRUCE ARNE SHERWOOD is assistant director and co-chairman of the System Software Group at the Computer-Based Education Research Laboratory and professor of physics and of linguistics at the University of Illinois at Urbana-Champaign. His research areas include Plato system software, computer-based physics teaching, multilingual text-to-speech synthesis, and problems of international communication. He received a BS in engineering science from Purdue, studied physics at Padua under a Fulbright scholarship grant, and received a Ph.D. in physics from the University of Chicago.

Spanish text. Moreover, the children were hearing their native language, so that conditions were more favorable than in our own work using a less sophisticated phonemic synthesizer with standard Esperanto text and with students learning a new language.

Few-Word Speech Recognition

There now exist speaker-dependent word recognition systems [7] that can identify isolated words drawn from a vocabulary of a few hundred words. Accuracy is fairly high, but the system must be trained with a particular individual's voice. There also exist systems that can handle a vocabulary of 10 to 20 isolated words independent of the speaker. Marie-Thérèse Giorgetti-Mas of the University of Nancy (France) has used a system of the latter kind to drill beginning students on Esperanto pronunciation [8].

For each drill item, the instructor records

the correct pronunciation of a word and the most common mispronunciations. For example, since a French student might mispronounce "glaso" (glass) as "glazo," the instructor will record both forms. When using the system, the student sees the word "glaso" on the terminal's screen and must pronounce it. The word recognition system has just a two-word vocabulary ("glaso" and "glazo") to compare with the student's pronunciation. The student can be told the pronunciation is correct or that the "s" is wrong. Of course, if the student's pronunciation is too far from either "glaso" or "glazo," the system can only ask the student to try again, since it doesn't really evaluate the individual sounds. The key to the success of this approach is that in this drill context there are only a few words to be compared, and such comparisons can be relatively independent of the speaker.

In tests of the system at the University of Paderborn (West Germany), the scheme

Esperanto: An Overview

LUDWIG ZAMENHOF of Poland introduced the constructed language Esperanto in 1887 to provide a politically neutral and easily learnable language for speakers of many different nationalities.

Since the Second World War, English has been widely used as a language of international communication. However, in recent years political and economic changes have led to increasing technical and other uses of such languages as Japanese, French, German, Russian, Spanish and Arabic. The number of official languages in the European Community, the United Nations, and other international organizations has been growing. These trends have led to renewed interest in the role Esperanto can play in improving international communications.

One expression of this new interest is the linguistics course International Communication and Constructed Languages at the University of Illinois, in which synthetic speech has been used experimentally in the teaching of Esperanto.

Esperanto is written in a simple phonetic manner. Each letter always has just one pronunciation, and stress is always on the next to the last vowel. There are only five vowels, as in Spanish (English has a dozen vowel sounds, which are often difficult for foreigners: "beat," "bit," "bait," "bet," "bat," "Bob," "bought," "boat," "book," "boot," "but," and "Bert"). The word "Esperanto" is pronounced "Es-peh-RAHN-toe." The phonetic writing, regular stress, uniform rhythm, and comparatively simple system of sounds combine to make it easy to synthesize good-quality speech from standard text.

The vocabulary roots are drawn mainly from European languages, but the way in which words are formed and modified is very different from the European languages. An unusual and distinctive aspect of the language is that there are endings that unambiguously identify the part of speech. Adjectives end in "a"; nouns end in "o"; adverbs end in "e"; and verbs end in "is" (past tense), "as" (present), "os" (future), "us" (conditional), "u" (imperative), or "i" (infinitive). For example, from the roots "rapid" and "help" many related words can be formed in a completely regular way:

rapida = fast	helpa = helpful
rapido = speed	helpo = (a) help
rapide = rapidly	helpe = helpfully
rapidis = hurried	helpis = helped
rapidas = hurries	helpas = helps
rapidos = will hurry	helpos = will help
rapidus = would hurry	helpus = would help
rapidu = hurry!	helpu = help!
rapidi = to hurry	helpi = to help

Verbs are not conjugated, and there are no irregular verbs or irregular plurals (basically, there are no irregularities of any kind). Although part-of-speech information has not yet been used in the Esperanto text-to-speech algorithm, the endings provide grammatical information that could be used to improve the phrasing of what is already rather good synthetic speech.

In addition, there is a powerful set of affixes for making new words in a regular way. Take "-id-" as an example:

urso = bear	ursido = bear cub
bovo = cow	bovido = calf
hundo = dog	hundido = puppy
kato = cat	katido = kitten
	ido = offspring

Because of the regularity and simplicity illustrated here, Esperanto can be mastered to a higher level than is feasible with national languages. That makes the language a more efficient and accurate communications medium, as well as a more equitable one. Studies indicate that one year of college-level Esperanto study is roughly equivalent to four years of a national language such as French or German.

Main centers of Esperanto activity are West and East Europe, Japan, and China. As an example, China publishes an excellent monthly magazine in Esperanto and broadcasts 30-minute shortwave radio programs in Esperanto four times a day. Many new Esperanto courses have started recently in China.

There is a sizable and interesting literature in Esperanto, not only translations from national literatures, including the lesser known ones such as Finnish and Bulgarian, but also works written originally in Esperanto. The technical literature is not extensive, but in recent years activity in the computer field has been increasing, in the form of both international conferences and new publications. For example, the computer science association of Hungary is organizing an international computer conference for December 1982 in Budapest, at which Esperanto will be the working language.

A useful textbook that is widely available is *Teach Yourself Esperanto* by John Cresswell and John Hartley, with the companion volume *Esperanto Dictionary* by John Wells, both published by Hodder and Stoughton (London) and David McKay (New York). Further information can be obtained from the Esperanto League for North America, P.O. Box 1129, El Cerrito, Calif. 94530, or from the parent organization, Universala Esperanto-Asocio, Nieuwe Binnenweg 176, 3015 BJ Rotterdam, the Netherlands.

worked well with 20 different students, both adults and children, male and female, and with the instructor's recordings made by a woman. Performance can be improved if the instructor records each correct and incorrect word several times, to give the system a wider tolerance of pronunciations for matching purposes. Similarly, it can be useful to have several different instructors record versions of the words.

System Integration

In order to use text-to-speech synthesis easily and well, a number of system integration considerations should be kept in mind. Among these are computer language design to accommodate speech output, the ability to terminate a message, and the capability to synchronize speech and screen displays.

In the Tutor programming language [9] the basic display command is "write," and to display the text "La bela kato" ("The beautiful cat") on the screen the statement is "write La bela kato." There now exists an exactly parallel command, "say," and to make the Votrax synthesizer attached to the terminal speak to the student, one simply uses the statement "say La bela kato."

The Plato editing program provides interactive tools for composing line graphics, "write" statements (to create them in the context of a complete screen display) and "say" statements (so that the author can hear the sentence and edit it in terms of changing punctuation or emphasis). The author can use punctuation marks to affect phrasing and intonation, and words written all in uppercase are spoken with increased emphasis (e.g., KATO), making it possible to give some variety to the synthetic speech.

In order to present a data-base drill item on the screen the program statement is "write <a, mess, len>" where the special brackets indicate that what is to be written on the screen is not fixed but variable text. The "a" indicates "alphabetic" (as opposed to purely numeric), "mess" is a variable holding the character-string message and "len" specifies the number of letters in the text to be displayed. In a completely similar way, to speak this text one can use the statement "say, <a, mess, len>." Fixed and variable text can be combined. In the word dictation drill one generates "nat semee, semi" by a statement of the form:

say nat <a, stu, sl>, <a, mess, len>

where the student's response is placed in a location "stu" with "sl" number of letters. A preceding "saylang" command must appear in the program to specify the language of the later "say" commands (Esperanto, Spanish, World English Spelling, or International Phonetic Alphabet), until another "saylang" command is encountered.

Other System Considerations

It is important to be able to terminate an audio message before it is finished. This capability is needed in order to jump ahead to the next interaction if desired or to cut off a message that the student recognizes is redundant. Moreover, users at computer terminals easily tolerate repetitive displays but not repetitive audio, because the ear can't ignore things the way the eye can. In many cases it is appropriate simply to cut off any current message if another message is received by the synthesizer.

Telesensory Speech Systems has recently announced a text-to-speech synthesizer with a feature that will be extremely useful in system integration. To make it possible to coordinate speech and screen displays accurately, text sent to the synthesizer can contain numbered markers. When all the speech up to the marker has been produced, the number corresponding to that marker is sent back to the host. Thus the host can synchronize other activities with the speech output. 

REFERENCES

1. **B. A. Sherwood**, "New Technology Provides Computer Voices for Education," *Speech Technology* 1, no. 1 (Fall 1981): 24-29.
2. **B. A. Sherwood and C. C. Cheng**, "A Linguistics Course on International Communication and Constructed Languages," *Studies in the Linguistic Sciences* 10 (1980): 189-201. ERIC document 182995.
3. **J. N. Sherwood**, "Plato Esperanto Materials," *Studies in Language Learning* 3 (1981): 123-128.
4. **B. A. Sherwood**, "Fast Text-to-Speech Algorithms for Esperanto, Spanish, Italian, Russian, and English," *International Journal of Man-Machine Studies* 10 (1978): 669-692.
5. **B. A. Sherwood**, "Speech Synthesis applied to Language Teaching," *Studies in Language Learning* 3 (1981): 171-181.
6. **B. A. Sherwood**, "Spertoj pri sintezo de Esperanta parolado" [Experiences with synthesis of Esperanto speech] in *Lingvo-Kibernetiko* [Cybernetics of language], ed. Frank, Yashovardhan, and Frank-Böhringer (Tübingen, West Germany: Gunter Narr Verlag, 1982), pp. 63-74.
7. **J. M. Baker**, "How to Achieve Recognition: A Tutorial/Status Report on Automatic Speech Recognition," *Speech Technology* 1, no. 1 (Fall 1981): 30-43.
8. **M.-T. Giorgetti-Mas**, "Automata skribado de parolata lingvo" [Automatic writing of spoken language], in *Lingvo-Kibernetiko* [Cybernetics of language], ed. Frank, Yashovardhan, and Frank-Böhringer (Tübingen, West Germany: Gunter Narr Verlag, 1982), pp. 58-62.
9. **B. A. Sherwood**, "The TUTOR Language" (Minneapolis, Minn.: Control Data Education Company, 1977).

FOR FURTHER INFORMATION

Contact Prof. Bruce Arne Sherwood or Judith N. Sherwood, 252 Eng. Res. Lab, 103 S. Mathews, Urbana, Ill. 61801; (217) 333-6210.

DRAGON SYSTEMS

Welcomes your
inquiries into Speech Technology
Applications and R&D



"It pays to be recognized"

DRAGON SYSTEMS, INC.
173 Highland St.
West Newton, MA 02165
(617) 527-0372

FOR INFORMATION CIRCLE 14 ON READER CARD

World English Spelling (WES)

a	fat	o	on (between aa and au)
aa	faather (father)	oe	toe
ae	Mae	oi	oil
ar	far	oo	too
au	taut	or	for
b	but	ou	out
ch	chum	p	pet
d	dig	r	run
e	set	s	set
ee	see	sh	shed
er	gather	t	tin
f	fat	th	this
g	gum	thh	thhing (thing)
h	hat	u	up
i	in	ue	hue
ie	tie	ur	fur (a long "er")
j	jam	uu	buuk (book)
k	kit	v	van
l	let	w	win
m	met	wh	when
n	net	y	yes
ng	sing	z	zoo
nk	sink	zh	vizhun (vision)

Optional "synonomous" spellings can be used for even more naturalness: x for ks, qu for kw, oy for oi, ow for ou, ea for ee, dg for j, au for aw, and ai or ei for ae. When r is preceded by ae (or ai or ei), the vowel should be treated as "e" (as in "thaer"). When r is preceded by ee (or ea), the vowel should be treated as "i" (as in "beer"). When c is followed by e, i, or y, treat it as s: otherwise treat it as k (except for ch, of course).

One must separate ambiguous combinations with a period, as in "mis.hap" (to prevent the "sh" from having its usual sound). WES is designed in such a way that this is rarely necessary. Capitalize the beginning of stressed syllables:

Forscor and Seven Yearz uGoe, our Faatherz Braut Forthh
on this Continent u Noo Naeshn, cunSeevd in Liberty,
and Dedicatet too thu propuZishun that Aul Men ar
creeAeted Eequul.

Fig. 8. A phonetic spelling scheme for English.

TYPE as in

a fat
 aa faather (father)
 ae Mae
 ar far
 au taut
 b but
 ch chum
 d dig
 e set
 ee see
 er gather
 f fat
 g gum
 h hat
 i in
 ie tie
 j jam
 k kit
 l let
 m met
 n net
 ng sing
 nk sink

TYPE as in

o on (between aa & au)
 oe toe
 oi oil
 oo too
 or for
 ou out
 p pet
 r run
 s set
 sh shed
 t tin
 th this
 thh thhing (thing)
 u up
 ue hue
 ur fur (long "er")
 uu buuk (book)
 v van
 w win
 wh when
 y yes
 z zoo
 zh vizhun (vision)

HELP for more info

When necessary,
 use a period to
 separate letters,
 as in "mis.hap".

At the end of a
 phrase, use a
 comma or period
 for a short or
 long pause.

Capitalize the
 start of stressed
 syllables, as in
 "uPeel" (appeal).

Tiep u Sentens, uezing thu World Inglish Speling shoen uBuy.

>

World English Spelling (WES) is a scheme for spelling English the way it sounds. Note that "long vowels" end in "silent e": ae in Mae, ee in see, ie in tie, oe in toe, and ue in hue. Type a number to hear a sample:

1) for Guud Speech, Kapituliez Strest or imPortent Silublz. The comma makes a short pause and a rise in the pitch of the voice. Capitalize the important syllables. (";" can be used to make a short pause with no pitch rise.)

2) Short.hand iz u Kwik Wae too Riet. "Shorthand is a quick way to write." The period in short.hand prevents it sounding like shor thand.

3) ie wont too Goe. ie Wont too goe. Ie wont too goe. Emphasis is determined by capitalization. Periods make a long pause and a fall in the pitch of the voice. (":" can be used for a short pause and falling pitch.)

4) too Bee or Naat too Bee, THAT iz thu Kweschun. Capitalize an entire syllable for extra emphasis.

5) did yoo See it? whut Wuz it? Question marks make a long pause and a rise in the pitch of the voice (but pitch FALLS if the question starts with "wh" words, such as whut, when, whie).

6) 8+9 = 17. \$23.45 Numbers, expressions, and prices can be spoken.

Here are a few hints that you may find helpful.

- 1) The word "the" is spelled "thu" before a consonant, "thi" before a vowel, and "thee" when specially emphasized:

thu Bug. thi Ant. Thee best. THEE best.

- 2) The word "there" is spelled "the.r" or "thaer", NOT "ther", which would sound like the end of "gather". Similarly for where (whaer), air (aer), care (kaer), etc.

- 3) Listen for "s" or "z" at the end of a plural word:

Bets. Bedz. Kits. Kidz.

While WES is a complete system for spelling English, you may if you wish use the following "synonyms" for some of the WES forms:

ai = WES ae (maid)

aw = WES au (saw)

c before e, i, or y = WES s (cell, cinder, fancy)

c before other letters (not h) = WES k (cat)

dg = WES j (budget)

ea = WES ee (lean)

ei = WES ae (rein)

ow = WES ou (cow)

oy = WES oi (boy)

qu = WES kw (quiz)

x = WES ks (tax)

thu Maid saw Fancy Cats.

thu Budget iz Lean.

Rein thu Cow.

Tax thu Boy with u Quiz.

Special vowels

If you type "aeaeaeae", you will find that you don't simply get a sustained "ae" sound: it wavers. Similarly, "oeoeoeoe" and "oooooooo" waver. This effect is due to the fact that these vowels are actually diphthongs in English.

To make extremely long, sustained sounds, or to use "pure" vowels in foreign words, use

ay for ae
oa for oe
eu for oo

These give just the beginning parts of the English diphthongs. A very long "ae" can be typed as "ayayayay", and so on.

Automatic lengthening

In WES, vowels are automatically lengthened somewhat when stressed, and also when followed by one of the voiced consonants b, d, g, j, v, or z. This helps make "sad" sound different from "sat".

"Errors"

Illegal characters or too long a sentence will produce "ee" sounds to indicate the error.

May 27, 1983

Bruce Sherwood
University of Illinois at Urbana-Champaign
Computer-Based Education Research Laboratory
103 South Mathews Avenue
Urbana, IL 61801-2977

Dear Mr. Sherwood:

Thank you for taking the time to talk with me and to send the reprints. I have made cppyies of the package for the people at Atari who will be working with the speech peripheral. Thanks again for your help.

Yours very truly,

Harry B. Stewart